На правах рукописи

ЗАЙЦЕВ АЛЕКСЕЙ АЛЕКСЕЕВИЧ

МЕТОДЫ ПОСТРОЕНИЯ РЕГРЕССИОННЫХ МОДЕЛЕЙ РАЗНОРОДНЫХ ИСТОЧНИКОВ ДАННЫХ ДЛЯ ИНДУСТРИАЛЬНОЙ ИНЖЕНЕРИИ

05.13.18 —
Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2017

Работа выполнена в ФГБУ науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук

Научный руководитель:

Бурнаев Евгений Владимирович,
кандидат физико-математичеких наук, доцент,
заведующий лабораторией № 10 Интеллектуального анализа данных и предсказательного моделирования Института проблем передачи информации им. А.А. Харкевича РАН

Официальные оппоненты:

Назин Александр Викторович,
доктор физико-математических наук, профессор, ведущий, научный сотрудник, Института проблем управления РАН

Красоткина Ольга Вячеславовна,
кандидат физико-математических наук,
доцент кафедры математических методов прогнозирования факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова

Ведущая организация: Федеральное государственное бюджетное учреждение Национальный исследовательский центр "Курчатовский институт"

Защита состоится «___» _____ 2017 г. в _____ на заседании диссертационного совета Д 002.077.05 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича РАН, расположенном по адресу: 127051, г. Москва, Большой Каретный переулок, д.19 стр. 1.

С диссертацией можно ознакомиться в библиотеке ФГБУ науки Института проблем передачи информации им. А.А. Харкевича РАН и на сайте www.iitp.ru.

Автореферат разослан «___» _____ 2017 г.

Ученый секретарь
диссертационного совета,
д.ф.-м.н. И. И. Цитович

Общая характеристика работы

Актуальность темы. В индустриальной инженерии одной из основных задач является задача проектирования изделия, характеристики которого удовлетворяют заданным требованиям. Часто в задачах индустриальной инженерии применяют подход, основанный на использовании быстро вычислимой регрессионной модели, построенной по выборке пар «параметры изделия (входной вектор, точка) — его характеристики (выходной вектор)», где характеристики изделия получаются в результате ресурсоемкого численного моделирования или натурных экспериментов.

В задачах индустриальной инженерии используемые данные могут быть разными по точности и стоимости получения, разнородными: часть данных может быть порождена источником данных высокой точности, а другая часть — источником данных низкой точности, при этом ресурсоемкость использования источника данных высокой точности обычно существенно выше ресурсоемкости использования источника данных низкой точности. Например, в задаче построения модели зависимости подъемной силы крыла самолета от его формы данные высокой точности могут быть получены из экспериментов в аэродинамической трубе, а данные низкой точности — из расчетов с помощью численного моделирования. При наличии разнородных источников данных можно выбирать для каких изделий использовать источник данных высокой точности, а для каких — низкой, чтобы для заданного общего ресурсного ограничения построить по полученным данным как можно более точную регрессионную модель.

Не существует универсального алгоритма для построения регрессионных моделей. Часто применяют метод, основанный на предположении о том, что моделируемая функция есть реализация гауссовского процесса. Такой метод называют регрессией на основе гауссовских процессов. Он широко используется для постро-

ения нелинейных регрессионных моделей, в том числе и по выборкам разнородных данных. Для однородных данных исследования регрессии на основе гауссовских процессов проводились в работах А.Н. Колмогорова, М. Штайна, А. Ван Дер Ваарта, и других, минимаксная ошибка — ошибка для наилучшей аппроксимации для наихудшей целевой функции заданной гладкости — была получена в работе Г.К. Голубева и Е.А. Крымовой. Однако, для разнородных данных существующие результаты либо опираются на эвристики, либо получены в предположениях, не позволяющих использовать такие результаты на практике. Часто эффективный план экспериментов для разнородных источников данных таков, что требуется использование чрезмерных вычислительных ресурсов для построения регрессионной модели. Однако, на сегодняшний день вычислительно эффективные подходы к регрессии на основе гауссовских процессов разработаны только для однородных данных. Таким образом, **актуальны** разработка вычислительно эффективных методов построения регрессионных моделей разнородных данных на основе гауссовских процессов, проведение исследования таких моделей данных и разработка метода выбора эффективного плана экспериментов для разнородных источников данных для подхода на основе гауссовских процессов в условиях заданного ресурсного ограничения.

Объектом исследования являются регрессионные модели индустриальной инженерии на основе гауссовских процессов, параметры которых оцениваются по разнородным данным. **Предметом** исследования являются методы построения регрессионных моделей разнородных данных и метод выбора эффективного плана экспериментов, предназначенного для построения таких моделей.

Целями данной работы является разработка вычислительно эффективных методов построения регрессионных моделей разнородных данных, оценка качества таких регрессионных моделей,

и разработка методов выбора эффективного плана экспериментов для таких моделей.

Поставленные цели определили следующие **задачи** исследования:

1. Разработать вычислительно эффективные методы построения регрессионных моделей разнородных данных, которые учитывают типичные особенности таких данных, и создать их программную реализацию.
2. Получить оценку качества регрессионных моделей данных на основе гауссовских процессов.
3. Разработать метод выбора эффективного плана экспериментов — соотношения между размерами выборок разнородных данных при заданном ресурсном ограничении, которое максимизирует качество получаемой регрессионной модели.

Научная новизна работы состоит в том, что в ней впервые были получены следующие результаты:

1. Разработан новый метод построения регрессионных моделей на основе гауссовских процессов по выборкам разнородных данных, основанный на численных методах для низкоранговой аппроксимации.
2. В многомерном случае получены минимаксные ошибки интерполяции для моделей нелинейной регрессии на основе гауссовских процессов, построенных по выборкам как однородных, так и разнородных данных.
3. Разработан новый метод выбора соотношения размеров выборок разнородных данных, минимизирующего минимаксную ошибку интерполяции при заданном ресурсном ограничении.

Теоретическая и практическая значимость представленной диссертационной работы определяется строгостью полученных математических результатов и широким использованием рассмотренных методов для моделирования по выборкам разнородных дан-

ных. Предложенные в работе методы используются для решения прикладных задач, возникающих в инженерной практике.

Общая методика исследования. Для решения поставленных задач в работе используются методы математической статистики, теории случайных процессов, аппарата анализа Фурье, статистической теории машинного обучения, матричной алгебры.

Основные положения, выносимые на защиту:

1. Разработанный метод построения нелинейных регрессионных моделей для выборок разнородных данных на основе низкоранговой аппроксимации имеет трудоемкость $O(\phi(n)^2 n)$ вместо $O(n^3)$ для стандартного подхода. Значение $\phi(n)$ обычно выбирают порядка $\min(c, n)$, где c — константа, задаваемая требованием к качеству модели.
2. Полученная теоретическая оценка качества регрессионной модели многомерных нелинейных зависимостей, в том числе в случае наличия разнородных источников данных, позволяет определить целесообразность использования разнородных источников данных.
3. Разработанный метод выбора соотношения между размерами выборок разнородных данных является теоретически оптимальным и обеспечивает высокое качество регрессионных моделей на практике.
4. Разработанные методы вошли в состав программного комплекса, предназначенного для решения задач анализа данных в индустриальной инженерии.
5. С помощью разработанного программного комплекса решен ряд задач индустриальной инженерии.

Достоверность изложенных в работе результатов определяется использованием корректных математических методов, основанных на хорошо изученных подходах из теории математической статистики; результатами проведенных численных экспериментов;

согласованностью полученных результатов с ранее известными; а также успешным использованием предложенных подходов для решения реальных задач индустриальной инженерии.

Апробация работы. Результаты диссертации докладывались и обсуждались на следующих конференциях: международная конференция молодых ученых «Информационные Технологии и Системы» (2012, Петрозаводск; 2016, Репино), 9-ая Международная конференция «Интеллектуализация обработки информации» (2012, Будва, Черногория), Conference on Structural Inference in Statistics (2013, Потсдам, Германия), конференции 5th Symposium on Conformal and Probabilistic Prediction with Applications (2016, Мадрид, Испания). Также результаты работы обсуждались на семинарах лаборатории структурных методов анализа данных в предсказательном моделировании МФТИ (2013, 2015, 2016, Москва), «Математические модели информационных технологий» НИУ ВШЭ (2015, Москва), отдела Интеллектуальных систем ВЦ РАН (2015, Москва), «Байесовские методы машинного обучения» ВМК МГУ (2015, Москва), И. Оселедца Сколтеха (2016, Москва); различных лабораторий ИППИ РАН (2016, Москва).

Публикации. Основные результаты по теме диссертации изложены в 7 печатных работах, из которых 6 [1–6] изданы в журналах, рекомендованных ВАК.

Личный вклад автора. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Подготовка к публикации выносимых на защиту результатов проводилась совместно с соавторами.

В цикле работ [1, 2, 3] постановки задач принадлежат соавторам, доказательство результатов получено лично диссертантом. Идеи некоторых вычислительных экспериментов принадлежат Е.В. Бурнаеву, постановка экспериментов и анализ их результатов были сделаны автором.

В работе [4] основные результаты получены автором, вычислительные эксперименты проведены автором. Е.В. Бурнаевым предложены постановки задач.

Структура и объем диссертации. Диссертация состоит из введения, 5 глав, заключения и библиографии. Общий объем диссертации 148 стр., включая 46 рисунков. Библиография включает 114 наименований.

Благодарности. Автор благодарен своему научному руководителю Е.В. Бурнаеву за постановки задач, плодотворные обсуждения, помощь в подготовке работ по теме диссертации к публикации, постоянную поддержку. Автор также благодарен Г.К. Голубеву за профессиональную экспертизу полученных результатов и ценные обсуждения.

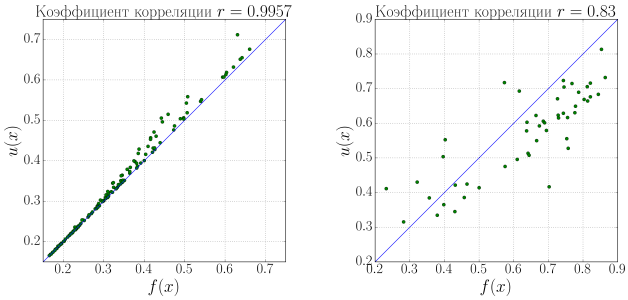
Содержание работы

Во **введении** показана актуальность диссертационной работы, аргументирована цель исследования, продемонстрирована важность рассматриваемой задачи с точки зрения фундаментальной науки и прикладных приложений полученных результатов.

Первая глава посвящена месту рассматриваемых задач в индустриальной инженерии и их формальному математическому описанию.

Суррогатного моделирование — подход к математическому моделированию, основанный на построении регрессионных моделей по выборкам данных. В большом числе приложений данные могут быть получены из разнородных источников, например, в результате использования и натурных экспериментов, и результатов численного моделирования. Сравнение значений целевых характеристик крыла самолета и вращающегося диска, полученных с помощью разнородных источников данных, приведено на Рисунке 1.

В исследовании рассматриваются модели регрессии на основе гауссовских процессов, в том числе и для разнородных источников



(а) Прочность вращающегося диска

(б) Подъемная сила крыла самолета

Рисунок 1 – Сравнение результатов, полученных для одних и тех же входных векторов с использованием более точного ($u(\mathbf{x})$) и менее точного ($f(x)$) источников данных для двух разных задач индустриальной инженерии

данных (в литературе традиционно говорят о регрессии на основе гауссовских процессов в том числе для многомерных входных векторов, хотя более правильным русским названием будет «регрессия на основе гауссовских случайных полей»). Определим используемую модель данных разной точности[1]

$$u(\mathbf{x}) = \rho f(\mathbf{x}) + g(\mathbf{x}), \tag{1}$$

здесь функция $u(\mathbf{x})$ соответствует источнику данных высокой точности, а $f(\mathbf{x})$ соответствует источнику данных низкой точности, ρ — константа, определяющая насколько близки источники данных разной точности $u(\mathbf{x})$ и $f(\mathbf{x})$, а $g(\mathbf{x})$ — поправка, описывающая отличие источников данных разной точности.

Предполагается, что $f(\mathbf{x})$ и $g(\mathbf{x})$ — реализации двух независимых стационарных гауссовских процессов на $\mathbb{R}^d$ со спектральными плотностями $F(\boldsymbol{\omega})$ и $G(\boldsymbol{\omega})$ соответственно. Мы наблюдаем значения $u(\mathbf{x})$ и $f(\mathbf{x})$ в некотором наборе точек и строим регрессионную мо-

[1] M.C. Kennedy и A. O'Hagan. "Predicting the output from a complex computer code when fast approximations are available". В: Biometrika 87.1 (2000), с. 1—13.

дель (аппроксимацию) $\widetilde{u}(\mathbf{x})$ источника данных высокой точности.

Пусть мы наблюдаем $u(\mathbf{x})$ на $D^u = D_H$, а $f(\mathbf{x})$ — на $D^f = D_{\frac{H}{m}}, m \in \mathbb{Z}^+$, где план эксперимента $D_H = \{\mathbf{x} : \mathbf{x} = H\mathbf{k}, \mathbf{k} \in \mathbb{Z}^d\}$, H — диагональная матрица с элементами на диагонали $h_1, \ldots, h_d$. Определим ошибку интерполяции для $\widetilde{u}(\mathbf{x})$ как:

$$\sigma^2_{H,m}(\widetilde{u}, F, G, \rho) \stackrel{\text{def}}{=} \frac{1}{\mu(\Omega_H)} \int_{\Omega_H} \mathbb{E}\left[\widetilde{u}(\mathbf{x}) - u(\mathbf{x})\right]^2 d\mathbf{x}, \tag{2}$$

где $\Omega_H = [0, h_1] \times \ldots \times [0, h_d]$, а мера $\mu(\Omega_H) = \prod_{i=1}^d h_i$.

В диссертации рассматривается для заданных наблюдений ошибка интерполяции $\sigma^2_{H,m}(\widetilde{u}, F, G, \rho)$ в случае известных спектральных плотностей и в минимаксном случае — когда спектральные плотности процессов $f(\cdot)$ и $g(\cdot)$ неизвестны, но удовлетворяют следующим неравенствам:

$$\mathbb{E}\sum_{i=1}^d \left[\frac{\partial f(\mathbf{x})}{\partial x_i}\right]^2 \le L_f, \ \mathbb{E}\sum_{i=1}^d \left[\frac{\partial g(\mathbf{x})}{\partial x_i}\right]^2 \le L_g.$$

Вторая глава посвящена задаче построения нелинейной регрессионной модели на основе гауссовских процессов. Пусть задана выборка $\mathbf{S} = (D^u, \mathbf{u}) = \{\mathbf{x}_i, u_i = u(\mathbf{x}_i)\}_{i=1}^n$ размера n, состоящая из точек $\mathbf{x}_i \in \mathbb{X} \subseteq \mathbb{R}^d$ и значений целевой функции $u_i = u(\mathbf{x}_i)$ в этих точках. Задача состоит в построении модели $\widetilde{u}(\mathbf{x})$ целевой функции $u(\mathbf{x})$ с использованием выборки $\mathbf{S}$.

Мы предполагаем, что $u(\mathbf{x})$ является реализацией гауссовского процесса. В этой главе будем предполагать, что тренд в точках известен, и его вычли из данных, то есть среднее процесса равно нулю, а ковариационная функция процесса имеет вид

$$c(\mathbf{x}, \mathbf{x}') = \operatorname{cov}(u(\mathbf{x}), u(\mathbf{x}')) = \mathbb{E}\left(u(\mathbf{x}) - \mathbb{E}u(\mathbf{x})\right)\left(u(\mathbf{x}') - \mathbb{E}u(\mathbf{x}')\right).$$

Тогда совместное распределение вектора $\mathbf{u}$ — многомерное нормальное, $\mathbf{u} \propto \mathcal{N}(\mathbf{0}, \mathbf{C})$, где $\mathbf{C} = \{c(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ — ковариационная матрица.

Условное распределение $u(\mathbf{x})$ при заданной выборке $\mathbf{S}$ в точке

$\mathbf{x} \in \mathbb{R}^d$ тоже будет нормальным:

$$u(\mathbf{x})|\mathbf{S} \propto \mathcal{N}(\mu(\mathbf{x}), v^2(\mathbf{x})).$$

Выражения для условного среднего $\mu(\mathbf{x})$ и дисперсии $v^2(\mathbf{x})$ выписываются аналитически. Среднее $\mu(\mathbf{x})$ используется как регрессионная модель $\widetilde{u}(\mathbf{x})$, а дисперсия $v^2(\mathbf{x})$ — как оценка неопределенности модели.

Используют параметрическое предположение о ковариационной функции: $c(\mathbf{x}, \mathbf{x}') \in \{c_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}'), \boldsymbol{\theta} \in \Theta \subseteq I\!R^p\}$. Тогда для построения модели нужно оценить параметры ковариационной функции $\boldsymbol{\theta}$. В качестве $\boldsymbol{\theta}$ обычно используется оценку максимума правдоподобия или байесовскую оценку параметров модели регрессии на основе гауссовских процессов.

Третья глава посвящена вычислительно эффективным методам построения регрессионной модели разнородных данных. В случае наличия источников разнородных данных с использованием модели (1) получаем формулы, аналогичные приведенным во второй главе. Пусть мы наблюдаем значения $u(\mathbf{x})$ и $f(\mathbf{x})$ зашумленные белым шумом, причем дисперсии шума для них соответственно σ_u^2 и σ_f^2.

Обозначим $D = \left(D^f, D^u\right)^\mathsf{T}$, $\mathbf{y} = (\mathbf{f}, \mathbf{u})^\mathsf{T}$. Тогда для модели источника данных высокой точности в точках D_*:

$$\widetilde{\mathbf{u}}(D_*) = \mathbf{C}(D_*, D)\mathbf{C}(D, D)^{-1}\mathbf{y},$$

где

$$\mathbf{C}(D_*, D) = \begin{pmatrix} \rho\mathbf{C}_f(D_*, D^f) \\ \rho^2\mathbf{C}_f(D_*, D^u) + \mathbf{C}_g(D_*, D^u) \end{pmatrix},$$

$$\mathbf{C}(D, D) = \begin{pmatrix} \mathbf{C}_f(D^f, D^f) & \rho\mathbf{C}_f(D^f, D^u) \\ \rho\mathbf{C}_f(D^u, D^f) & \rho^2\mathbf{C}_f(D^u, D^u) + \mathbf{C}_g(D^u, D^u) \end{pmatrix},$$

$\mathbf{C}_f(D_a, D_b)$, $\mathbf{C}_g(D_a, D_b)$ — матрицы попарных ковариаций гауссовских процессов $f(\mathbf{x})$ и $g(\mathbf{x})$ в точках из D_a и D_b соответственно.

Условная матрица ковариации имеет вид:

$$v^2\left(D_*\right) = \rho^2 \mathbf{C}_f(D_*, D_*) + \mathbf{C}_g(D_*, D_*) - \mathbf{C}(D_*, D)\mathbf{C}(D, D)^{-1}\mathbf{C}(D_*, D)^{\mathsf{T}}.$$

Если размеры выборок превышают несколько тысяч точек, непосредственное использование такой модели становится невозможным, так как вычислительная сложность алгоритма оценки параметров модели — $O(n^3)$, где $n = n_u + n_f$, $n_u = |D^u|$, $n_f = |D^f|$.

В диссертационном исследовании предложен вычислительно эффективный метод построения регрессионной модели разнородных данных. Метод основан на использовании низкоранговой аппроксимации Нистрема матриц $\mathbf{C}(D_*, D)$, $\mathbf{C}(D, D)$ и $\mathbf{C}(D_*, D_*)$. Представленные результаты являются обобщением результатов работы Л. Фостера[2] на случай разнородных данных.

Выберем *базовые* точки из исходной выборки: $\mathbf{S}_1 = (D_1, \mathbf{y}^1)$, $D_1 = (D_1^u, D_1^u)^{\mathsf{T}}, \mathbf{y}^1 = (\mathbf{f}(D_1^u), \mathbf{u}(D_1^u))^{\mathsf{T}}$, причем размеры $n_f^1 = |D_1^u|$, $n_u^1 = |D_1^u|$.

Положим

$$\mathbf{C}_{11} = \begin{pmatrix} \mathbf{C}_f(D_1^u, D_1^u) & \rho\mathbf{C}_f(D_1^u, D_1^u) \\ \rho\mathbf{C}_f(D_1^u, D_1^u) & \rho^2\mathbf{C}_f(D_1^u, D_1^u) + \mathbf{C}_g(D_1^u, D_1^u) \end{pmatrix},$$

$$\mathbf{C}_1 = \begin{pmatrix} \mathbf{C}_f(D_1^u, D^f) & \rho\mathbf{C}_f(D_1^u, D^u) \\ \rho\mathbf{C}_f(D_1^u, D^f) & \rho^2\mathbf{C}_f(D_1^u, D^u) + \mathbf{C}_g(D_1^u, D^u) \end{pmatrix},$$

$$\mathbf{C}_1^* = \begin{pmatrix} \rho\mathbf{C}_f(D_*, D_1^u) \\ \rho^2\mathbf{C}_f(D_*, D_1^u) + \mathbf{C}_g(D_*, D_1^u). \end{pmatrix}$$

Определим также $M = \begin{pmatrix} \frac{1}{\sigma_f} I_{n_f} & 0 \\ 0 & \frac{1}{\sqrt{\rho^2\sigma_f^2 + \sigma_u^2}} I_{n_u} \end{pmatrix}$, $V = M\mathbf{C}_1 V_{11}^{-T}$, V_{11} — разложение Холецкого $\mathbf{C}_{11}$.

Утверждение 1. *Для приближения ковариационных матриц* $\widehat{\mathbf{C}}(D_*, D) = \mathbf{C}_1^*\mathbf{C}_{11}^{-1}\mathbf{C}_1^{\mathsf{T}}$, $\widehat{\mathbf{C}} = M^{-2} + \mathbf{C}_1\mathbf{C}_{11}^{-1}\mathbf{C}_1^{\mathsf{T}}$ *низкоранговая ап-*

[2] L. Foster и др. “Stable and efficient Gaussian process calculations”. В: The Journal of Machine Learning Research 10 (2009), с. 857—882

проксимация апостериорного среднего имеет вид:

$$\widetilde{u}(D_*) = \mathbf{C}_1^* V_{11}(I_{n_1} + V^\mathsf{T} V)^{-1} V^\mathsf{T} \mathbf{y},$$

а низкоранговая аппроксимация апостериорной дисперсии имеет вид: $v^2(D_*) = \mathbf{C}(D_*, D_*) - \mathbf{C}_1^* V_{11}^{-1}(I_{n_1} + V^\mathsf{T} V)^{-1}(V^\mathsf{T} V) V_{11}^{-\mathsf{T}} (\mathbf{C}_1^*)^\mathsf{T}$. *Вычислительная сложность использования таких аппроксимаций —* $O(n_1^2 n)$.

При этом для использованного в работе метода выбора базовых точек выполнена оценка качества, аналогичная приведенной в работе Кумара [3] а именно с вероятностью $1 - \delta$

$$\|\mathbf{C}(D_*, D_*) - \widehat{\mathbf{C}}(D_*, D_*)\|_2 \le \|\mathbf{C}(D_*, D_*) - \widehat{\mathbf{C}}_{n_1}(D_*, D_*)\|_2 + \Delta_{**},$$
$$\|\mathbf{C}(D_*, D) - \widehat{\mathbf{C}}(D_*, D)\|_2 \le \|\mathbf{C}(D_*, D) - \widehat{\mathbf{C}}_{n_1}(D_*, D)\|_2 + \Delta_*,$$

где Δ_* и Δ_{**} порядка $O\left(\frac{n}{\sqrt{n_1}}\right)\left(1 + O\left(\sqrt{\log \frac{1}{\delta}}\right)\right)^{\frac{1}{2}}$, $\|\cdot\|_2$ — l_2 матричная норма, а $\widehat{\mathbf{C}}_{n_1}(D_*, D_*)$ — наилучшая в смысле l_2 нормы аппроксимация ранга $n_1 = n_f^1 + n_u^1$.

Четвертая глава содержит основные теоретические результаты работы.

Рассмотрим $u(\mathbf{x})$ — реализацию стационарного гауссовского процесса с ковариационной функцией $c(\mathbf{x}_0, \mathbf{x}) = c(\mathbf{x} - \mathbf{x}_0)$ на $\mathbb{R}^d$. Спектральная плотность $F(\boldsymbol{\omega})$ — преобразование Фурье $c(\cdot)$: $F(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} e^{2\pi \mathrm{i} \boldsymbol{\omega}^\mathrm{T} \mathbf{x}} c(\mathbf{x}) d\mathbf{x}$.

Пусть известны значения реализации $u(\cdot)$ на бесконечной анизотропной сетке $D_H = \{\mathbf{x} : \mathbf{x} = H\mathbf{k}, \mathbf{k} \in \mathbb{Z}^d\}$, где H — диагональная матрица. Рассматривается ошибка $\sigma_H^2(\widetilde{u}, F)$ аппроксимации $\widetilde{u}(\mathbf{x})$ функции $u(\mathbf{x})$ на $\Omega_H = [0, h_1] \times \ldots [0, h_d]$:

$$\sigma_H^2(\widetilde{u}, F) \stackrel{\text{def}}{=} \frac{1}{\mu(\Omega_H)} \int_{\Omega_H} \mathbb{E}\left[\widetilde{u}(\mathbf{x}) - u(\mathbf{x})\right]^2 d\mathbf{x}.$$

В реальных задачах истинная ковариационная функция неизвестна, поэтому рассматривают минимаксную ошибку интерполя-

[3] Kumar, S. и др. “Sampling methods for the Nyström method”. В: The Journal of Machine Learning Research 13 (2012), с. 981—1006

ции. Определим класс спектральных плотностей $F(\boldsymbol{\omega})$ гауссовских процессов $\mathcal{F}(L)$:

$$\mathcal{F}(L) \stackrel{\text{def}}{=} \left\{ F : \mathbb{E} \sum_{i=1}^{d} \left(\frac{\partial u_F(\mathbf{x})}{\partial x_i} \right)^2 \leq L, \mathbf{x} \in \mathbb{R}^d \right\},$$

где $u(\mathbf{x}) = u_F(\mathbf{x})$ — гауссовский процесс со спектральной плотностью $F(\boldsymbol{\omega})$.

Минимаксную ошибку интерполяции определим как:

$$R^H(L) \stackrel{\text{def}}{=} \inf_{\widetilde{u}} \sup_{F \in \mathcal{F}(L)} \sigma_H^2(\widetilde{u}, F).$$

Теорема 1. *Для класса спектральных плотностей $\mathcal{F}(L)$ на $\mathbb{R}^d$ и наблюдений $u(\mathbf{x})$ на D_H минимаксная ошибка интерполяции имеет вид*

$$R^H(L) = \frac{L}{2\pi^2} \max_{i \in \{1,\ldots,d\}} h_i^2.$$

Более того, $\widetilde{u}(\mathbf{x})$, минимизирующая $\sup_{F \in \mathcal{F}(L)} \sigma_H^2(\widetilde{u}, F)$, имеет вид $\widetilde{u}(\mathbf{x}) = \mu(\Omega_H) \sum_{\mathbf{x}' \in D_H} K(\mathbf{x} - \mathbf{x}') u(\mathbf{x}')$, где $K(\mathbf{x})$ — симметричное ядро, преобразование Фурье которого $\widehat{K}(\boldsymbol{\omega})$ равно:

$$\widehat{K}(\boldsymbol{\omega}) = \begin{cases} 1 - \sqrt{\sum_{i=1}^d h_i^2 \omega_i^2}, & \sum_{i=1}^d h_i^2 \omega_i^2 \leq 1, \\ 0, & \sum_{i=1}^d h_i^2 \omega_i^2 > 1. \end{cases}$$

Перейдем к исследованию модели данных разной точности (1).

Теорема 2. *Для наблюдений $u(\mathbf{x})$ на D_H и $f(\mathbf{x})$ на $D_{\frac{H}{m}}$ минимум ошибки интерполяции (2) равен:*

$$\sigma_{H,m}^2(\widetilde{u}, F, G, \rho) = \sigma_H^2(\widetilde{g}, G) + \rho^2 \sigma_{\frac{H}{m}}^2(\widetilde{f}, F),$$

где $\widetilde{g}(\mathbf{x})$ и $\widetilde{f}(\mathbf{x})$ минимизируют соответственно $\sigma_H^2(\widetilde{g}, G)$ и $\sigma_{\frac{H}{m}}^2(\widetilde{f}, F)$.

Получим теперь минимаксную ошибку интерполяции. Пусть спектральные плотности процессов $f(\cdot)$ и $g(\cdot)$ принадлежат $\mathcal{F}(L_f)$ и $\mathcal{F}(L_g)$ соответственно. Для ясности ограничим изложение случаем $H = hI$. Определим

$$R^{h,m}(L_f, L_g) \stackrel{\text{def}}{=} \inf_{\widetilde{u}} \sup_{\substack{F \in \mathcal{F}(L_f), \\ G \in \mathcal{F}(L_g)}} \sigma_{H=hI,m}^2(\widetilde{u}, F, G, \rho). \tag{3}$$

Теорема 3. *Минимаксная ошибка интерполяции* (3) *для гауссовского процесса* $u(\mathbf{x}) = \rho f(\mathbf{x}) + g(\mathbf{x})$ *и наблюдений* $u(\mathbf{x})$ *на* D_H *и* $f(\mathbf{x})$ *на* $D_{\frac{H}{m}}$ *имеет вид*

$$R^{h,m}(L_f, L_g) = \rho^2 \frac{L_f}{2} \left(\frac{h}{m\pi}\right)^2 + \frac{L_g}{2} \left(\frac{h}{\pi}\right)^2. \tag{4}$$

Получим теперь оптимальное соотношение между размерами выборок данных разной точности в следующей постановке:

- Стоимость вычисления $u(\mathbf{x})$ равняется $\mathrm{w} > 1$, а $f(\mathbf{x})$ — 1;
- Общий бюджет B ограничен количеством точек в гиперкубе со стороной 1.

Таким образом, $B = \mathrm{w}\frac{1}{h^d} + \delta\frac{1}{h^d}$, где $\delta = m^d$ — отношение между размерами выборок разной точности. Тогда минимум минимаксной ошибки интерполяции (4) по δ, h для бюджета B имеет вид

$$R^{h,(\delta^*)^{\frac{1}{d}}}(L_f, L_g, \rho) = \rho^2 \frac{L_f}{2} \left(\frac{\mathrm{w} + \delta^*}{\pi B \delta^*}\right)^{\frac{2}{d}} + \frac{L_g}{2} \left(\frac{\mathrm{w} + \delta^*}{\pi B}\right)^{\frac{2}{d}},$$

а оптимальное соотношение между размерами выборок $\delta^* = \left(\frac{L_f}{L_g}\mathrm{w}\rho^2\right)^{\frac{d}{d+2}}$. Если мы можем использовать только источник данных высокой точности, то для бюджета B получаем следующую минимаксную ошибку интерполяции

$$R^h(L_f, L_g, \rho) = \rho^2 \frac{L_f}{2} \left(\frac{\mathrm{w}}{\pi B}\right)^{\frac{2}{d}} + \frac{L_g}{2} \left(\frac{\mathrm{w}}{\pi B}\right)^{\frac{2}{d}}.$$

Обозначим $R_2 = R^{h,(\delta^*)^{\frac{1}{d}}}(L_f, L_g, \rho)$ и $R_1 = R^h(L_f, L_g, \rho)$. Тогда значения L_f, L_g, ρ и w определяют соотношение между R_1 и R_2: если $R_2 < R_1$ использование разнородных данных позволяет уменьшить минимаксную ошибку интерполяции для регрессионной модели $\widetilde{u}(\mathbf{x})$ в рамках заданного общего бюджета. Естественно описывать соотношение между R_2 и R_1 в зависимости от коэффициента корреляции r между источниками данных разной точности $u(\mathbf{x})$ и $f(\mathbf{x})$. Если данные нормализованы и $\mathbb{E}f^2(\mathbf{x}) = \mathbb{E}g^2(\mathbf{x}) = 1$, то

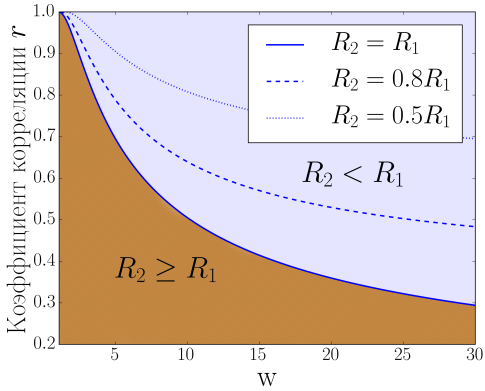


Рисунок 2 – Области, в которых $R_2 \geq R_1$ и $R_2 < R_1$ в зависимости от значений r и w, $L_f = 2$, $L_g = 1$, $d = 1$.

коэффициент корреляции:

$$r = \mathrm{corr}(u(\mathbf{x}), f(\mathbf{x})) = \frac{\mathbb{E}u(\mathbf{x})f(\mathbf{x})}{\sqrt{\mathbb{E}u^2(\mathbf{x})}\sqrt{\mathbb{E}f^2(\mathbf{x})}} = \frac{1}{\sqrt{1+\frac{1}{\rho^2}}}.$$

На Рисунке 2 приведены области, в которых $R_2 \geq R_1$ и $R_2 < R_1$ в зависимости от значений r и w.

В исследовании предложен Алгоритм 1 для оценки оптимального отношения между размерами выборок разной точности δ^*, основанный на полученных теоретических результатах.

Алгоритм 1 – Создание планов экспериментов D^f и D^u

Вход: Коэффициент корреляции r, бюджет B, относительная стоимость использования источника данных высокой точности w

1: $\rho^2 \leftarrow 1/(1-\frac{1}{r^2})$, $\delta^* \leftarrow (\mathrm{w}\rho^2)^{\frac{d}{d+2}}$
2: $n_f \leftarrow \frac{B\delta^*}{\mathrm{w}+\delta^*}$, $n_u \leftarrow \frac{B}{\mathrm{w}+\delta^*}$
3: Получить планы экспериментов D^f, D^u, $D^u \subseteq D^f$ размером $|D^f| = n_f$, $|D^u| = n_u$ из случайного равномерного распределения на $[0,1]^d$.
4: **Вернуть** D^f, D^u.

Пятая глава содержит примеры использования предложенных подходов к моделированию разнородных данных.

В условиях заданного ресурсного ограничения сравнивается работа предложенного в исследовании метода для выбора соотношения между размерами выборок данных разной точности **Minimax** с используемыми в литературе методами: **High** — используются только данные высокой точности, **EqSize** — размеры выборок высокой и низкой точности равны, **EqBudget** — доли бюджета, использованные источниками данных разной точности, равны.

В Таблице 1 приведены ошибки RRMS — среднеквадратичные ошибки аппроксимации, деленные на среднеквадратичную ошибку аппроксимации константой. Рассматривались задачи построения моделей зависимостей характеристик индустриального изделия от его параметров для крыла самолета (**Airfoil**), С-образного пресса (**Press**) и вращающегося диска (**Disk**). Часто с помощью подхода **Minimax** удается получить наилучшую или близкую к наилучшей по качеству регрессионную модель.

Таблица 1 – RRMS ошибки усредненные по 20 запускам скользящего контроля.

Задача	Выход	High	EqSize	EqBudget	Minimax
Airfoil	1	0.546	0.594	0.539	**0.5221**
Airfoil	2	**0.120**	0.142	0.130	0.138
Press	1	0.559	0.601	0.358	**0.277**
Press	2	0.443	0.491	0.271	**0.176**
Disk	1	0.299	0.340	**0.192**	0.193
Disk	2	0.446	0.457	**0.299**	**0.299**

Предложенные в исследовании подходы, реализованные соискателем в среде `Matlab`, были переписаны на `C++` и вошли в состав разработанного в компании DATADVANCE программного комплекса, использующегося в платформе pSeven, предназначенной для автоматизации инженерного проектирования, оптимизации и анализа

данных. Характерный пример использования разработанного комплекса для построения регрессионных моделей разнородных данных приведен на Рисунке 3.

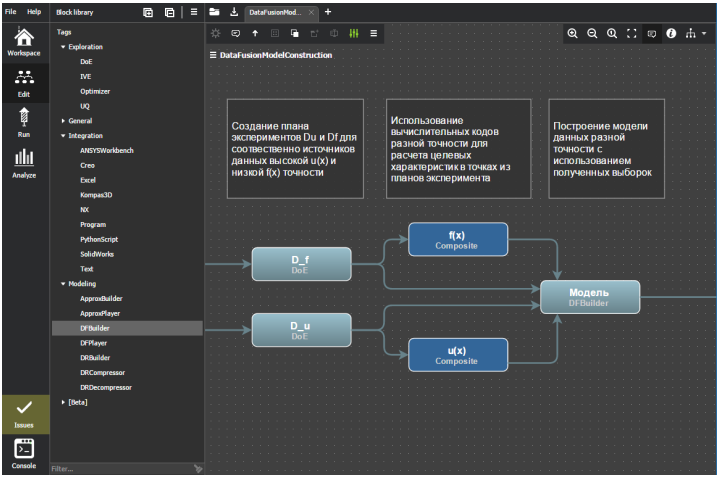


Рисунок 3 – Процесс построения регрессионной модели разнородных данных в pSeven: крайний правый блок `Модель` принимает на вход разнородные данные и, используя разработанные в исследовании подходы, строит регрессионную модель.

В **заключении** приведены основные результаты работы:

1. Разработан вычислительно эффективный алгоритм построения нелинейных регрессионных моделей разнородных данных.
2. Получена оценка качества многомерной регрессионной модели однородных и разнородных данных.
3. Получен метод выбора эффективного соотношения между размерами выборок данных разной точности.
4. Разработанные методы вошли в состав программного комплекса, предназначенного для решения задач анализа данных в индустриальной инженерии.
5. С помощью разработанного программного комплекса решен ряд задач индустриальной инженерии.

Публикации автора по теме диссертации

1. Зайцев А. А., Бурнаев Е. В., Спокойный В. Г. Свойства байесовской оценки параметров регрессии на основе гауссовских процессов // Фундаментальная и прикладная математика. 2013. Т. 18, № 2. С. 53–65.

2. Зайцев А. А., Бурнаев Е. В., Спокойный В. Г. Свойства апостериорного распределения модели зависимости на основе гауссовских случайных полей // Автоматика и телемеханика. 2013, Т. 74, № 10, С. 55–67.

3. Бурнаев Е. В., Зайцев А. А., Спокойный В. Г. Теорема Бернштейна–фон Мизеса для регрессии на основе гауссовских процессов // Успехи математических наук. 2013. Т. 68, № 5. С. 179–180.

4. Burnaev E.V., Zaytsev A.A. Surrogate modeling of multifidelity data for large samples // Journal of Communications Technology and Electronics, 2015. V. 60, № 12. P. 1348–1355.

5. Zaytsev A. Variable fidelity regression using low fidelity function blackbox and sparsification // Lecture Notes in Computer Science, 2016. V. 9653, P. 147–164.

6. Zaytsev A. Reliable surrogate modeling of engineering data with more than two levels of fidelity // IEEE ICMAE conference, 2016. V. 9653, P. 341–345.

7. Зайцев А. Ошибка интерполяции для регрессии на основе данных разной точности // ИТИС 2016, 40-я междисциплинарная школа-конференция, 15-20 сентября, 2016. С. 289–294.