Информационные процессы, Том 19, № 3, 2019, стр. 271–283

= МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ =

Алгоритмические основы предиктивной аналитики в задачах индустриального проектирования

Е.В. Бурнаев

Сколковский институт науки и технологий, Москва, Россия
Институт проблем передачи информации, Российская академия наук, Москва, Россия
e-mail: e.burnaev@skoltech.ru

Поступила в редколлегию 01.07.2019

Аннотация—Рассматривается задача построения предсказательных моделей (суррогатных моделей) для решения задач индустриальной предиктивной аналитики. Автор, проанализировав потребности индустриальных приложений, сформулировал ряд новых математических и алгоритмических задач и разработал соответствующие методы моделирования по данным.

Ключевые слова: предиктивная (предсказательная) аналитика, индустриальная инженерия, машинное обучение, предсказательное техническое обслуживание, предсказание поломок, многомерная регрессия, снижение размерности

1. ВВЕДЕНИЕ

В последние годы методы предсказательной аналитики стали необходимым инструментом рабочего процесса в аэрокосмической и нефтегазовой промышленности, биомедицинских исследованиях, финансовом секторе и т.п.

Для решения важных практических задач обслуживания оборудования на основе автоматического контроля его состояния, обеспечения безопасности сложных технических и информационных систем (самолетов, судов, ракет, ядерных электростанций, различных интернет-сервисов, и т.д.), автоматического контроля качества выпускаемой продукции, предсказания естественных катастрофических явлений (землетрясения, цунами, и т.д.), мониторинга в биомедицине, финансовой и социальной сферах необходимо построение моделей многокомпонентных динамических систем для прогнозирования их поведения, обнаружения аномалий и разладок.

При изучении таких систем для получения информации об их функционировании проводят натурные эксперименты и моделирование на основе “физики процессов”, в качестве которой в зависимости от конкретной прикладной области могут выступать физические, а также экономические, биологические, химические или другие законы. В свою очередь в современной предсказательной аналитике используются так называемые метамодели (суррогатные модели), которые “обучаются” по множеству прототипов входных и выходных данных (результатов натурных и/или вычислительных экспериментов, проведенных с различными объектами рассматриваемого

класса) и фактически имитируют (заменяют) как источники получения данных, основанные на некоторой исходной модели, так и сами модели, созданные на основе изучения “физики процессов”.

Например, в процессе индустриального проектирования происходит сравнение различных конструкторских решений, касающихся структуры и большого числа параметров проектируемого объекта, его поведения в тех или иных условиях окружающей среды. Численные расчёты возникающих при классическом моделировании дифференциальных уравнений имеют значительную вычислительную трудоёмкость. Это существенно сокращает возможности использования таких моделей, особенно на стадии предварительного проектирования, когда рассматривается большое количество разнообразных конструкторских решений, выбрать среди которых вручную наиболее эффективное даже опытный инженер обычно не в состоянии. Для примера, при проектировании современного самолета рассматривается до 100 000 вариантов компоновки, а для анализа прочностных характеристик силовых элементов самолета необходимо провести несколько миллионов численных экспериментов при разных значениях десятков параметров, определяющих конструкцию силового элемента. Таким образом, скорость проектирования и оптимизации изделия существенно ограничивается. В свою очередь, использование суррогатных моделей, построенных по относительно небольшому объёму данных, позволяет многократно сократить время на выявление более эффективных конструкторских решений за счет высокой вычислительной эффективности прогноза их характеристик.

Другим важным примером использования суррогатных моделей на практике является прогностическое управление производственными системами и экономическими процессами, обычно описываемыми большим количеством параметров. Действительно, например, эффективность процесса экстракции (диффузии) сахара-песка зависит от формы свекольной стружки, её качества, значения сахаристости, pH, температуры подаваемой воды, от распределения температуры внутри диффузора, и т.д. При этом, для минимизации затрат и снижения потерь необходим выбор оптимальных значений параметров. Очевидно, что даже опытный технолог не в состоянии уследить за всем многообразием управляющих воздействий, условий внешней среды, их взаимозависимостями и влиянием на эффективность производства. Соответственно, суррогатные модели, построенные по данным, накопленным в результате эксплуатации производственных установок, и лежащие в основе рекомендательных систем, существенно облегчают технологу решение его основных задач по выбору управляющих параметров производственного процесса (своего рода “второе мнение”) и позволяют снизить затраты и повысить качество продукции.

Рассмотрим ещё одно важное приложение суррогатных моделей —- предсказательное обслуживание. Методы предсказательного технического обслуживания используются для скорейшего обнаружения аномалий и существенных изменений (разладок) в работе механизмов и сложных технических систем. Основной целью этого подхода к обслуживанию является корректировка технического состояния или полная замена механизмов до того, как выявленные изменения станут критичными для работы отдельных компонент или системы в целом. Именно построение суррогатных

моделей позволяет выявить взаимосвязи между параметрами системы в нормальном режиме работы, и в дальнейшем, путем сравнения выявленных взаимосвязей с текущими результатами телеметрии, снимаемой с датчиков в режиме реального времени, обнаружить аномалии и разладки в работе системы.

Таким образом, можно выделить три основных типа задач индустриальной предиктивной аналитики, где необходимо использование суррогатных моделей:

Тип I: задачи индустриального проектирования. Основная цель — сокращение времени проектирования;
Тип II: управление на основе предиктивных моделей. Основные цели —

1. Оптимизация производственных процессов. Например, требуется сократить потребление сырья при производстве продукции определенного типа с заданным качеством;
2. Управление параметрами непрерывного производства. Например, требуется сократить потребление энергии, повысить эффективность производства;

Тип III: обнаружение аномалий и прогнозирование поломок. Основные цели —

1. Прогнозирование качества продукции на ранних стадиях производства и снижение объёма брака;
2. Предсказательное обслуживание.

Использование стандартных методов статистики и машинного обучения (Machine Learning, ML), таких как линейная регрессия, нейронные сети, градиентный бустинг на деревьях решений, и т.п. зачастую не позволяет построить суррогатную модель, имеющую достаточную для решения задач предсказательной аналитики точность. Причина в том, что задачи обработки индустриальных, биомедицинских и экономических данных обладают целым рядом особенностей и на их решения зачастую накладываются специфические требования, что существенно ограничивает использование стандартных методов общего назначения.

Цель работы — в описании формальных постановок задач анализа данных, решение которых необходимо при построении суррогатных моделей для индустриального проектирования, а также в обзоре новых методов, эффективно учитывающих особенности приложений и удовлетворяющих требованиям этих постановок задач.

2. СУРРОГАТНЫЕ МОДЕЛИ В ИНДУСТРИАЛЬНОМ ПРОЕКТИРОВАНИИ

Одним из основных препятствий на пути использования т.н. инженерной оптимизации, то есть применения специализированных алгоритмов оптимизации целевых функций, задаваемых вычислительными и/или натурными экспериментами, является длительное время выполнения симуляций, в которых один расчет может занимать от минут до часов и даже суток, а также отсутствие информации о градиенте оптимизируемого функционала и высокая стоимость натурных экспериментов.

Например, в компании Airbus при оптимизации запаса прочности самолета используется программа для расчета прочности композитного элемента самолета (стрингера), характеризуемой коэффициентами запаса прочности (Reserve Factors) в зависимости от геометрии конструкции, параметров материала и действующих сил (порядка 150 параметров). При снижении времени работы программы в 100 раз про-

гнозируется сокращение оптимизации структуры обшивки самолета с нескольких дней до нескольких часов (по оценкам компании Airbus) [22].

Основная идея "суррогатного моделирования" состоит в том, чтобы использовать быстрые аналитические модели (суррогатные модели), аппроксимирующие зависимости между параметрами, значения которых порождаются посредством компьютерного моделирования и/или дорогостоящего натурного эксперимента. Используя такого рода приближения можно проводить исследования типа "что-если", визуализировать зависимости между параметрами, проводить сравнительный анализ различных вариантов дизайна изделия, и т.п. Выбрав таким образом более эффективные варианты дизайна, их можно верифицировать, проведя соответствующие расчеты на основе компьютерного моделирования и/или натурного эксперимента, при необходимости пополнив набор расчетов новыми данными и перестроив аппроксимационную модель с учетом этих данных. Описанный итеративный подход называется суррогатной оптимизацией (surrogate optimization, SBO).

Несмотря на то, что идея такого подхода "лежит на поверхности", основные сложности возникают при описании формальных постановок задач и практической реализации соответствующих алгоритмов. Перечислим основные практические вопросы, для ответов на которые развиваются подобного рода инструменты:

– Как формулировать практические задачи индустриального проектирования в виде задач анализа данных?
– Как исследовать "физическую" модель в области изменения параметров проектирования, провести анализ "что-если"?
– Для каких комбинаций значений параметров дизайна изделия надо провести вычисление целевого функционала, чтобы с одной стороны минимизировать число таких вычислений, а с другой стороны построить как можно более точную суррогатную модель? Как оценить чувствительность значения целевого функционала к различным параметрам дизайна?
– Какие методы суррогатного моделирования использовать и как учитывать требования предметной области?
– Как использовать суррогатную модель для выбора новых улучшенных вариантов дизайна изделия: получение приемлемого решения или решения, улучшающего первоначальные характеристики объекта (проектирование "от прототипа")?
– Что делать, если в данных моделирования и/или натурного эксперимента присутствует численный шум и/или экспериментальный шум соответственно?

Последней, но не по значимости, причиной такой актуальности суррогатного моделирования является применение суррогатной оптимизации (SBO) в так называемом "автоматическом" ML (Auto ML). Действительно, эффективность моделей и алгоритмов ML в значительной степени зависит от специалистов по машинному обучению, которые выбирают подходящие архитектуры ML и их гиперпараметры. Как следствие, быстрый рост приложений машинного обучения создал спрос на "самонастраивающиеся" методы ML, использование которых не требует значительных экспертных знаний. Таким образом, рассматривая настройку гиперпараметров алгоритмов ML как задачу оптимизации черного ящика с дорогой в смысле времени вычисления и ресурсов, требуемых для этого, целевой функцией, мы можем использовать SBO для целей автоматизации ML.

3. ПОСТРОЕНИЕ СУРРОГАТНЫХ МОДЕЛЕЙ

Приведем постановки основных задач ML, которые необходимо решать при построении суррогатных моделей [7, 21, 23, 24].

Пусть $\mathbf{O} = \{O\}$ есть множество объектов рассматриваемого класса. Для каждого объекта $O \in \mathbf{O}$ имеется его цифровое описание $D = D(O)$ размерности d. Обозначим через $\mathbf{D} = \{D(O), O \in \mathbf{O}\}$ подмножество d–мерного евклидова пространства $\mathbb{R}^d$, состоящее из цифровых описаний объектов рассматриваемого класса.

Обозначим через $\mathbf{y} = \mathbf{y}(O) \in \mathbb{R}^q$ некоторую числовую (в общем случае, векторную) характеристику объекта O, описывающую свойство (поведение) объекта в некоторых условиях. Характеристика $\mathbf{y}$ может быть описана в виде функциональной зависимости $\mathbf{y} = f(\mathbf{x})$, $\mathbf{x} = (D, C)$, в которой переменная $D = D(O)$ описывает сам объект, а переменная C описывает условия функционирования объекта (параметры управления объектом, параметры внешней среды) и принадлежит множеству различных значений $\mathbf{C} \subset \mathbb{R}^c$. Таким образом, $\mathbf{x} \in \mathbf{X} = \mathbf{D} \times \mathbf{C} \subset \mathbb{R}^p$ $(p = d + c)$.

Например, аэродинамические характеристики самолета (коэффициенты сил, моментов, сопротивлений, распределения давлений и др.) в условиях крейсерского полета являются функцией, зависящей от формы поверхности самолета (переменная D) и параметров режима полета и управления (переменная C) (скорости, числа Рейнольдса, углов атаки, скольжения, установки горизонтального оперения, и др.).

Для вычисления значений функции $\mathbf{y} = f(\mathbf{x})$ могут использоваться различные методы (например, вычислительные или натурные эксперименты). Пусть M — некоторый метод, описывающий зависимость $\mathbf{y}$ от $\mathbf{x}$, тогда результат $f^M(\mathbf{x})$ вычисления значения $\mathbf{y}$ для входного вектора $\mathbf{x}$, задающего описание объекта и условия его функционирования, является приближением неизвестной функции $f(\mathbf{x})$, т.е. $f^M(\mathbf{x}) \approx f(\mathbf{x})$.

Рассмотрим основную задачу суррогатного моделирования – построение по данным $\mathbf{S}_n = \mathbf{S}_n(M) = \{(\mathbf{x}_i, \mathbf{y}_i = f^M(\mathbf{x}_i)), i = 1, 2, \ldots, n\}$, полученным с помощью модели $\mathbf{y} = f^M(\mathbf{x})$, $\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^p$ для значений $\mathbf{x}$ из множества $\mathbf{X}_n = \{\mathbf{x}_i, i = 1, 2, \ldots, n\} \subset \mathbf{X}$, эмпирической зависимости $\mathbf{y} = f_n(\mathbf{x}) \equiv f_n(\mathbf{x}|\mathbf{S}_n)$, аппроксимирующей f^M в области $\mathbf{X}$: $f_n(\mathbf{x}) \approx f^M(\mathbf{x})$ для всех $\mathbf{x} \in \mathbf{X}$.

Отметим, что аналитический вид аппроксимируемых функций, как правило, неизвестен. Во многих случаях функции $f^M(\mathbf{x})$ являются решениями сложных дифференциальных уравнений в частных производных, для которых неизвестны ни теоремы существования и единственности решения, ни теоремы о непрерывной зависимости решений от начальных условий. Поэтому знания о поведении аппроксимируемых функций, в том числе знания о виде зависимости функции от конкретных входных параметров, носят, как правило, только качественный характер.

В свою очередь функция $f_n(\mathbf{x})$ имеет явный аналитический вид, и поэтому суррогатная модель может иметь существенно более высокую вычислительную эффективность по сравнению с $f^M(\mathbf{x})$. Например, суррогатные модели для расчета аэродинамических характеристик, построенные по результатам, полученным с помощью CFD-модели, увеличили скорость вычислений более чем в 350 000 раз [8, 9].

Рассмотрим основные задачи, возникающие при построении и использовании суррогатных моделей. Самой очевидной является

Задача 1. Построение по $\mathbf{S}_n$ аппроксимирующей функции (аппроксиматора).

К этой задаче тесно примыкает

Задача 2. Оценивание точности суррогатной модели. Математически проблема сводится к задаче оценивания различных функционалов от величины ошибки $\delta(\mathbf{x}) = f_n(\mathbf{x}) - f^M(\mathbf{x})$, в том числе построения доверительных интервалов для $f^M(\mathbf{x})$.

Следующая проблема связана с тем, что для реальных задач размерность d цифрового описания D объекта может быть очень высока. Например, детальные описания геометрических объектов (кривых, поверхностей) или их компонентов в общем случае задаются набором $2D$- или $3D$-координат точек поверхности объекта, лежащих в выбранных узлах геометрического объекта. Другие точки объекта восстанавливаются, как правило, с использованием сплайнов. Такого рода детальные описания (например, $3D$-поверхности самолета) кривых и поверхностей используются в CAD-системах, компьютерной графике и других приложениях.

Таким образом, входные данные $\mathbf{x} = (D, C) \in \mathbf{X} \subset \mathbb{R}^p$ для "сложных" объектов имеют размерность порядка десятков или сотен чисел. В силу высокой размерности, множество $\mathbf{X}_n$ аргументов функции, для которых известны приближенные значения аппроксимируемой функции, является очень "разреженным".

Однако в приложениях множество $\mathbf{X}$ зачастую лежит, по крайней мере приближенно, на некотором многообразии (в общем случае, нелинейном) $\mathbf{X}^q \subset \mathbb{R}^p$, размерность q которого существенно меньше p. Следовательно, необходимо решать задачу аппроксимации лишь в окрестности многообразия $\mathbf{X}^q$.

Нахождение этого многообразия $\mathbf{X}^q$ очень важно и по другой причине. Для проведения вычислительных экспериментов с целью получения обучающего множества данных $\mathbf{S}_n$, необходимо генерировать входные данные $\mathbf{X}_n$ в окрестности многообразия $\mathbf{X}^q$ (в частности, необходимо "оставаться" вблизи многообразия $\mathbf{X}^q$ при генерации новых описаний объектов в процессе SBO).

Значит, при построении суррогатных моделей решаются две связанные задачи:

Задача 3. Построение многообразия меньшей размерности, аппроксимирующего данные. Математически, проблема сводится к задаче снижения размерности.

Задача 4. Генерация многомерных векторов вблизи построенного по данным аппроксимирующего многообразия меньшей размерности.

Построенная суррогатная модель $f^M(\mathbf{x}|\mathbf{S}_n)$, которая служит приближением к "истинной" функции $f(\mathbf{x})$, зависит как от исходной модели M, так и от данных $\mathbf{S}_n$.

Пусть уже имеется суррогатная модель $f_n^M(X)$ для некоторой исходной модели M, задающей приближение "истинной" функции $f(\mathbf{x})$. Рассмотрим новую исходную модель M^*, являющуюся более точным приближением (high fidelity model) той же "истинной" функции $f(\mathbf{x})$, чем модель M (low fidelity model). Для построения суррогатной модели $f_m^{M^*}(\mathbf{x})$, необходимо иметь результаты экспериментов $\mathbf{S}_m(M^*)$, проведенных с моделью M^*. В ситуации, когда эксперименты с моделью M^* являются

существенно более дорогостоящими (например, натурные эксперименты в аэродинамической трубе) по сравнению с моделью M (например, расчеты с помощью CFD-кода), т.е. $m \ll n$, возникает

Задача 5. Построение прогноза $f^{M^*}_{m,n}(\mathbf{x}) = \Phi\left(\mathbf{x}, f^M_n(\mathbf{x}) \middle| \mathbf{S}_m(M^*), \mathbf{S}_n(M)\right)$ значения $f^{M^*}(\mathbf{x})$ по известному значению $f^M_n(\mathbf{x})$ на основе результатов экспериментов $\mathbf{S}_m(M^*)$ и $\mathbf{S}_n(M)$, проведенных с моделями M^* и M соответственно.

Эффективное решение задачи 5 позволяет существенно сократить число экспериментов m с моделью M^* по сравнению с числом экспериментов, которые необходимо провести для достижения той же точности без учета уже построенной суррогатной модели $f^M_n(\mathbf{x})$ и соответствующей выборки $\mathbf{S}_n(M)$.

Модели, основанные на данных (прототипах), по своей сути могут гарантированно использоваться только для таких новых входных данных, которые подобны данным (прототипам) из множества $\mathbf{S}_n(M)$, с помощью которых была построена модель. Тем самым, для нового множества прототипов модель должна быть либо построена заново, либо перестроена (путем решения статистической задачи предсказания значений одной модели по значениям другой модели). Более того, очевидно, что от расположения точек обучающей выборки относительно особенностей (области разрывов, области большого градиента, и т.п.) аппроксимируемой модели значительным образом будет зависеть точность финальной суррогатной модели. Естественным образом возникает следующий адаптивный подход к построению суррогатных моделей и оптимизации на их основе:

1. Инициализация: проводятся вычислительные эксперименты с исходной физической модель $f(\mathbf{x})$ в точках $\mathbf{X}_n$, и в результате строится обучающая выборка $\mathbf{S}^{\text{init}}_n = \{\mathbf{x}_i, \mathbf{y}_i = f(\mathbf{x}_i)\}_{i=1}^n$;
2. По $\mathbf{S}_{\text{init}}$ строится функция $f_n(\mathbf{x})$: $f_n(\mathbf{x}) \approx f(\mathbf{x})$, $\mathbf{x} \in \mathbf{X}$;
3. Проводится поиск $\mathbf{x}_{\text{new}} \in \mathbf{X}$ и вычисление с исходной моделью f в точке $\mathbf{x}_{\text{new}}$, после чего выборка пополняется, то есть $\mathbf{S}_{n+1} = \mathbf{S}^{\text{init}}_n \cup \{\mathbf{x}_{\text{new}}, f(\mathbf{x}_{\text{new}})\}$;
4. Суррогатная модель $f_{n+1}(\mathbf{x})$ перестраивается по новой выборке;
5. Шаги $3-4$ повторяются;
6. Критерий завершения работы: по точности/времени.

Отметим, что в представленном выше алгоритме выбор новой точки $\mathbf{x}_{\text{new}} \in \mathbf{X}$ можно осуществлять исходя из разных конечных целей, а именно

1. Мы можем выбирать $\mathbf{x}_{\text{new}} \in \mathbf{X}$ так, чтобы добавление точки $\{\mathbf{x}_{\text{new}}, f(\mathbf{x}_{\text{new}})\}$ в выборку $\mathbf{S}^{\text{init}}_n$ позволяло при перестроении по выборке $\mathbf{S}_{n+1}$ суррогатной модели максимально увеличить её точность;
2. Другой вариант — выбирать $\mathbf{x}_{\text{new}} \in \mathbf{X}$ так, чтобы значение $\{\mathbf{x}_{\text{new}}, f(\mathbf{x}_{\text{new}})\}$ как можно лучше соответствовало экстремальному значению исходной модели $f(\mathbf{x})$.

При этом, в обоих случаях при выборе следующего значения требуется учитывать оценку точности построенной суррогатной модели. Таким образом, возникает ещё и

Задача 6. Разработка алгоритмов адаптивного выбора $\mathbf{x}_{\text{new}} \in \mathbf{X}$ для "активного" построения (active learning) суррогатной модели и SBO на её основе.

Так как для большинства указанных задач не существует эффективных универсальных процедур, то суррогатные модели создаются специалистами в предметной области и в анализе данных, знакомыми с различными математическими методами

и понимающими, как структура данных влияет на качество той или иной процедуры. Такой специалист может “вручную” выбрать или построить достаточно эффективную частную процедуру анализа данных, в том числе путем проведения сравнительных вычислительных экспериментов. Финальная суррогатная модель строится путем взаимосвязанного последовательного решения ряда частных задач анализа данных и проведения сравнительных вычислительных экспериментов.

Для возможности использования суррогатных моделей непосредственно в процессе проектирования имеется настоятельная необходимость создания программных средств автоматической генерации статистических процедур, входом которых будут массивы данных (обучающие выборки), а выходом будут являться программные модули, реализующие те или иные процедуры обработки данных (снижения размерности, аппроксимации и т.п.). Генераторы процедур в процессе своей работы имитируют деятельность исследователя, целенаправленно проводя вычислительные эксперименты с различными процедурами анализа данных и синтезируя наиболее эффективную процедуру для заданного множества данных. Тем самым, возникает

Задача 7. Разработка программных генераторов процедур машинного обучения, которые по заданной обучающей выборке автоматически создают программные модули, реализующие базовые составные части суррогатных моделей.

4. ХАРАКТЕРНЫЕ АСПЕКТЫ ЗАДАЧ МОДЕЛИРОВАНИЯ ПО ДАННЫМ

Перечислим характерные аспекты задач моделирования по данным, которые были выявлены в процессе анализа прикладных задач индустриальной инженерии.

Пространственная неоднородность. Моделируемая зависимость часто имеет разное поведение в различных областях пространства входных параметров, при этом большая часть вариабельности моделируемой зависимости обусловлена только небольшой долей компонент входного вектора. Построенные модели должны адекватно представлять эти неоднородности. Другой пример подобного типа состоит в анизотропных свойствах данных, в частности, дизайн эксперимента может быть достаточно специфическим. Например, во многих прикладных инженерных задачах часто используют факторный план эксперимента. В таких планах эксперимента переменные разбиваются на несколько групп, в каждой из которых переменные принимают значения из некоторого конечного множества. Группы переменных (и соответствующее множество значений) называют факторами; размер множества различных значений, которые могут принимать переменные из одного фактора, называется размером фактора, а сами значения — уровнями. Декартово произведение факторов образует полный факторный план эксперимента. Например, для различных значений числа Маха в аэродинамической трубе проводят натурные эксперименты при разных значениях угла атаки.

В реальных инженерных задачах зачастую факторный план эксперимента неполный, т.е. содержит пропуски. Это может быть связано, например, с тем, что не удалось провести измерение для некоторого набора входных данных в результате сбоя аппаратуры либо результаты компьютерных симуляций оказались нефизичными.

Использование факторного плана эксперимента (полного или неполного) приводит к тому, что размеры обучающих выборок получаются очень большими (размер выборки растет экспоненциально от количества входных переменных).

Предложен метод декомпозиции пространства входных параметров на области, в пределах которых моделируемая зависимость имеет регулярное поведение, см. [5,22]. Предложен метод моделирования нестационарной ковариационной функции гауссовского процесса на основе линейного разложения по словарю параметрических функций [3, 12, 20]. Использование такой ковариационной функции позволяет существенно повысить качество моделирования пространственно-неоднородных зависимостей, см. [17]. В [4, 6, 28] предложены подходы, которые позволяют эффективно работать с факторными планами экспериментов.

Локальная оценка точности метамоделей. Оптимизация построенных суррогатных моделей [2], например, для выбора оптимального дизайна проектируемого изделия, построение доверительного интервала прогнозируемой величины, и т.п. требует оценки локальной точности суррогатной модели.

Предложен новый подход к байесовской регуляризации параметров ковариационной функции гауссовского процесса, который позволяет повысить качество как самой модели, так и оценки её точности, см. [14, 17, 18]. Предложен новый метод построения регрессии на основе гауссовских процессов для структурированных выборок [6], который позволяет качественно оценивать точность в ряде задач с большим объемом выборки данных, см. также [27].

Гладкость суррогатной модели. В значительной части задач есть априорные ограничения на гладкость метамодели, которые должны быть учтены при её построении. Более того, в ряде случаев необходимо обеспечить возможность варьировать гладкость метамодели с целью нахождения её оптимального значения для конкретной прикладной задачи.

Разработаны методы подбора гладкости модели для алгоритма регрессии на основе гауссовских процессов, которые позволили сделать алгоритм адаптивным к широкому диапазону зависимостей различной гладкости, см. [5, 17].

Адаптивное планирование эксперимента. На практике, как правило, часто есть возможность выбирать, для каких значений входных описаний объектов произвести вычисления выходной характеристики. Необходимо оптимальным образом выбирать набор точек, в которых будет проводиться эксперимент, причем подбор новых точек должен идти совместно с построением суррогатной модели.

Предложена байесовская постановка задачи адаптивного планирования экспериментов, в рамках которой удалось обосновать как известные, так и предложить новые методы адаптивного планирования экспериментов, см. [15, 16].

Моделирование разнородных данных. В задачах индустриальной инженерии используемые данные могут быть разными по точности и стоимости получения, разнородными: часть данных может быть порождена источником данных высокой точности (например, CFD с плотной сеткой, натурные испытания), а другая часть — источником данных низкой точности (например, CFD с грубой сеткой, численные

симуляции), при этом ресурсоемкость использования источника данных высокой точности обычно существенно выше ресурсоемкости использования источника данных низкой точности. 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 метамодель. В такой постановке актуальными являются как задачи адаптивного планирования экспериментов и локальной оценки точности суррогатной модели, так и задачи изучения теоретических основ методов, позволяющих строить суррогатные модели в указанных условиях.

Предложен подход к планированию экспериментов для разнородных данных, основанный на гладкости разных источников данных, см. [31]. Разработан программный комплекс для построения метамоделей на основе разнородных данных, в том числе большого размера, см. [30]. Получены строгие результаты о теоретических основах разработанных методов, см. [19, 31].

Другие особенности. Указанные выше аспекты задач моделирования по данным — не единственные. Часто на практике

a) требуется прогнозировать ряд зависимых между собой величин, то есть суррогатная модель имеет многомерный выход. Типичный пример — прогнозирование давления на поверхности профиля крыла самолета. Для построения суррогатных моделей в этом случае предложены различные подходы на основе гауссовских процессов и нейронных сетей с многомерным выходом [3, 17].

b) между компонентами вектора $\mathbf{x}$, описывающего объект, имеются ряд зависимостей, обусловленных естественными ограничениями на значения этих компонент. Соответственно, это приводит к тому, что для объектов подобного типа их описание $\mathbf{x}$ приближенно лежит на некотором многообразии меньшей размерности. В таком случае, при построении суррогатной модели необходимо одновременно проводить оценку и самого многообразия [11, 25, 26]. Для этого предложены различные подходы, в т.ч. и учитывающие локальную структуру многообразия.

c) Выходная переменная $\mathbf{y}$ зависит не от всех компонент входного описания $\mathbf{x}$. Перед построением суррогатной модели требуется удалить те из компонент описания $\mathbf{x}$, зависимость от которых незначимая. Для этого разработаны специализированные методы оценки чувствительности и выделения признаков [15].

5. ЗАКЛЮЧЕНИЕ

В данной работе была рассмотрена задача построения предсказательных моделей (суррогатных моделей) для решения задач индустриального проектирования. Автор, проанализировав потребности индустриальных приложений, сформулировал ряд новых математических и алгоритмических задач и разработал соответствующие методы моделирования по данным.

В рамках проведенных научных исследований автором был получен ряд новых результатов. В частности, был предложен целый ряд ключевых концепций в разработанных методах предсказательной аналитики: подходы к иерархической байесовской регуляризации в регрессионных моделях, в т.ч. на основе гауссовских процес-

сов, и методы адаптивного планирования экспериментов; вычислительно эффективные алгоритмы построения регрессии на основе гауссовских процессов для данных со специальной структурой; непараметрические методы оценки точности прогноза суррогатной модели, и др.

Проведенные исследования в значительной степени были мотивированы потребностями задач индустриального проектирования. Это позволило использовать разработанные математические методы, реализованные компанией ООО "ДАТАДВАНС" (см. [13]) в виде алгоритмического ядра программной системы pSeven Core (aka MACROS library) суррогатного моделирования и инженерной оптимизации, для решения целого ряда прикладных инженерных задач. Разработанное алгоритмическое ядро получило сертификацию на финальном уровне Technology Readiness Level (TRL). Согласно официальному пресс-релизу компании Airbus, разработанные методы позволили сэкономить до 10% в некоторых важных этапах полного цикла проектирования новых моделей самолетов (см. [1]).

Использование разработанных алгоритмов и их дальнейшее развитие позволило успешно реализовать ряд инженерных приложений в авиакосмической промышленности [10], а также применить их для решения задач предсказательной аналитики в других прикладных областях, например,

– Моделирования прочностных характеристик элементов обшивки самолета A350-900 (компания Airbus), см. [22, 28];
– Моделирования аэродинамических характеристик суборбитального космического летательного аппарата (компания Astrium, см. [4]), вертолета (компания Airbus Helicopters, см. [29]), и др.

Е. Бурнаев выражает благодарность компании ООО Датадванс за предоставление данных и постановки задачи.

Исследование было частично поддержано грантом РФФИ 16-29-09649 офи_м.

СПИСОК ЛИТЕРАТУРЫ

1. Airbus uses datadvance's macros software for structural analysis of a350 xwb aircraft. https://www.aerospace-technology.com/news/newsairbus-uses-datadvances-macros-software-structural-analysis-a350-xwb-aircraft-4362144/.

2. A. Baranov, E. Burnaev, and et al. Optimising the active muon shield for the ship experiment at cern. J. of Physics: Conf. Ser., 934(1):012050, 2017.

3. M. Belyaev and E. Burnaev. Approximation of a multidimensional dependency based on a linear expansion in a dictionary of parametric functions. Informatics and its Applications, 7(3):114–125, 2013.

4. M. Belyaev, E. Burnaev, and et al. Building data fusion surrogate models for spacecraft aerodynamic problems with incomplete factorial design of experiments. Advanced Materials Research, 1016:405–412, 2014.

5. M. Belyaev, E. Burnaev, and et al. Gtapprox: Surrogate modeling for industrial design. Advances in Engineering Software, 102:29 – 39, 2016.

6. M. Belyaev, E. Burnaev, and Y. Kapushev. Computationally efficient algorithm for gaussian process regression in case of structured samples. Computational Mathematics and Mathematical Physics, 56(4):499–513, 2016.

7. A. Bernstein, E. Burnaev, and et al. About solution of some data analysis problems necessary when constructing adaptive surrogate models of complex objects (in russian). In Artificial Intelligence, volume 4, pages 40–48, Ukraine: Kyiv, 2008.

8. A. Bernstein, V. Vyshinskiy, A. Kuleshov, and Yu. Sviridenko. Fast method for aerodynamic calculations in engineering design problems (in russian). In Usage of artificial neural networks in applied aerodynamics problems. Proceedings of TSAGI named after Prof. N.E. Zhukovski, volume 2678. MOSCOW: TSAGI, 2008.

9. E. Burnaev. Fast method for calculation of aerodynamic characteristics of a passenger aircraft based on approximation of multidimensional dependencies (in russian). In Proceedings of 3rd All-Russia conference on control problems, pages 224–226, Moscow, Russia, 2008.

10. E. Burnaev and A. Bernstein. Methods of data analysis, predictive modelling and maintenance. In I. Uzhinsky, K. Sypalo, O. Aladyshev, and G. Rudenskiy, editors, Article (section 2.5) in a book "Advanced technologies for aerospace industry. Analytical overview" (In Russian), pages 117–127, Moscow, 2017. Nauka.

11. E. Burnaev and S. Chernova. On an iterative algorithm for calculating weighted principal components. J. of Comm. Tech. and Elect., 60(6):619–624, 2015.

12. E. Burnaev and P. Erofeev. The influence of parameter initialization on the training time and accuracy of a nonlinear regression model. J. of Comm. Tech. and Elect., 61(6):646–660, 2016.

13. E. Burnaev, F. Gubarev, S. Morozov, A. Prokhorov, and D. Khominich. PSE/MACROS: Software environment for process integration, data mining and design optimization (in russian). Iter-sectoral Information Service journal, (4):41–50, 2013.

14. E. Burnaev and I. Nazarov. Conformalized kernel ridge regression. In 2016 15th IEEE International Conference on Machine Learning and Applications, pages 45–52, Dec 2016.

15. E. Burnaev, I. Panin, and B. Sudret. Efficient design of experiments for sensitivity analysis based on polynomial chaos expansions. Annals of Math. and Art. Int., 81(1):187–207, 2017.

16. E. Burnaev and M. Panov. Adaptive design of experiments based on gaussian processes. In Statistical Learning and Data Sciences, pages 116–125. Springer, 2015.

17. E. Burnaev, M. Panov, and A. Zaytsev. Regression on the basis of nonstationary gaussian processes with bayesian regularization. J. of Comm. Tech. and Elect., 61(6):661–671, 2016.

18. E. Burnaev and V. Vovk. Efficiency of conformalized ridge regression. In COLT, volume 35, pages 605–622. PMLR, 2014.

19. E. Burnaev, A. Zaytsev, and V. Spokoiny. The Bernstein-von Mises theorem for regression based on Gaussian processes. Russ. Math. Surv., 68(5):954–956, 2013.

20. E. V. Burnaev and P. V. Prikhod'ko. On a method for constructing ensembles of regression models. Automation and Remote Control, 74(10):1630–1644, Oct 2013.

21. Alexander Forrester, Andreas Sobester, and Andy Keane. Engineering design via surrogate modelling: a practical guide. John Wiley & Sons, 2008.

22. S. Grihon, E. Burnaev, M. Belyaev, and P. Prikhodko. Surrogate Modeling of Stability Constraints for Optimization of Composite Structures, pages 359–391. Springer, 2013.

23. A. Kuleshov and A. Bernstein. Cognitive technologies in adaptive models of complex plants. In Keynote papers of 13th IFAC Symposium on Information Control Problems in Manufacturing (INCOM'09), pages 70–81, Moscow, Russia, 3–5 Jun 2009.

24. A. Kuleshov, A. Bernstein, and E. Burnaev. Adaptive models of complex systems based on data handling. In Proceedings of the 3rd International Conference on Inductive Modelling ICIM'2010, pages 64–71, Kyiv, Ukraine, 16–22 May 2010.

25. A. Kuleshov, A. Bernstein, and E. Burnaev. Kernel regression on manifold valued data. In Proceedings of IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA-2018), pages 120–129, 2018.

26. A. Kuleshov, A. Bernstein, and E. Burnaev. Manifold learning regression with non-stationary kernels. In Artificial Neural Networks in Pattern Recognition, pages 152–164. Springer, 2018.

27. M. Munkhoeva, Ye. Kapushev, E. Burnaev, and I. Oseledets. Quadrature-based features for kernel approximation. In NIPS, pages 9165–9174. 2018.

28. G. Sterling, P. Prikhodko, E. Burnaev, M. Belyaev, and S. Grihon. On approximation of reserve factors dependency on loads for composite stiffened panels. Advanced Materials Research, 1016:85–89, 2014.

29. A. Struzik, E. Burnaev, and P. Prikhodko. Surrogate models for helicopter loads problems. In Proc. of the 5th European Conf. for Aeronautics and Space Sciences, 2013.

30. A. Zaytsev and E. Burnaev. Large scale variable fidelity surrogate modeling. Annals of Mathematics and Artificial Intelligence, 81(1):167–186, Oct 2017.

31. A. Zaytsev and E. Burnaev. Minimax Approach to Variable Fidelity Data Interpolation. In Proceedings of the 20th AISTATS, volume 54 of PRML, pages 652–661. PMLR, 2017.

ALGORITHMIC FOUNDATIONS OF PREDICTIVE ANALYTICS IN INDUSTRIAL ENGINEERING DESIGN

Burnaev E.V.

Skolkovo Institute of Science and Technology, Moscow, Russia
Institute for Information Transmission Problems, Moscow, Russia

We consider the problem of constructing predictive models (surrogate models) to tackle challenges of industrial engineering design. The author analyzed the needs of industrial applications, formulated a number of new mathematical and algorithmic problems and developed appropriate methods of data modeling.

KEYWORDS: predictive analytics, industrial engineering, machine learning, predictive maintenance, failure prediction, multivariate regression, dimensionality reduction, surrogate modelling