Информационные процессы, Том 8, № 1, 2008, стр. 47–57.

МАТЕМАТИЧЕСКИЕ МОДЕЛИ, ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ

Снижение размерности при наличии предикатов[1]

А.В. Бернштейн*, А.П. Кулешов**

**Институт системного анализа РАН, Москва, Россия*
e-mail a.bernstein@cpt-ran.ru
***Институт проблем передач информации им. А.А. Харкевича РАН, Москва, Россия*
e-mail kuleshov@iitp.ru

Поступила в редколлегию 19.02.2008

Аннотация—Сформулирована задача снижения размерности многомерных данных при наличии "объясняющих переменных" (предикатов), возникающая при создании основанных на данных суррогатных адаптивных моделей. Такие модели предназначены для быстрого вычисления характеристик сложных объектов и основаны на когнитивных технологиях. Предложен подход к решению задачи и приведены результаты моделирования, иллюстрирующие эффективность этого подхода.

1. ВВЕДЕНИЕ

В процессе проектирования и создания сложных технических многокомпонентных систем рассматриваются и сравниваются различные технические решения, касающиеся структуры систем, механизмов их функционирования, выбора параметров и других элементов объекта. Для сравнения решений и выбора оптимального (рационального) решения создаются основанные на знаниях системы поддержки инженерных решений, в основе которых лежит широкий спектр различных когнитивных технологий. Для таких решений необходимо уметь сравнивать характеристики (свойства) объекта для различных вариантов его построения и в различных условиях функционирования. Ограниченные возможности проведения натурных и вычислительных экспериментов для получения значений характеристик для различных вариантов проектируемого объекта, а также низкая точность эвристических инженерных методов делают актуальной проблему создания технологий, основанных на упрощенных моделях [1, 2] и позволяющих в режиме реального времени проводить сравнение большого числа вариантов построения сложных технических объектов с обеспечением требуемой достоверности выводов.

Одной из таких востребованных технологий является когнитивная технология быстрых расчетов [3–6], позволяющая строить основанные на данных так называемые суррогатные адаптивные модели. Построенные модели фактически имитируют (заменяют) как источники получения данных об объекте, основанные на некоторой исходной (полноразмерной или упрощенной) модели, так и сами модели, созданные на основе изучения физических феноменов, описывающих процессы функционирования объектов.

Обе модели (исходная и суррогатная) должны иметь один и тот же набор входных и выходных данных, а результаты обоих моделей (для одних и тех же входных данных) должны быть близкими.

[1] Эта работа выполнена в рамках Проекта "Теоретические основы информационной технологии быстрых расчетов для основанных на знаниях систем поддержки инженерных решений" (руководитель А.П. Кулешов) Программы фундаментальных исследований Российской Академии Наук (Отделение нанотехнологий и информационных технологий) "Фундаментальные основы информационных технологий и систем" (Руководитель С.В. Емельянов).

Для создания адаптивных моделей необходимо решение ряда новых теоретических задач. Наряду с универсальными когнитивными технологиями, такими как технологии искусственного интеллекта, извлечения данных (Data Mining), моделирования и анализа данных и др., когнитивная технология быстрых расчетов включает в себя предметно-ориентированные компоненты, основанные на решении новых математических задач анализа и обработки многомерных данных (снижения размерности, аппроксимации зависимостей, оценивания точности и др.).

В статье приведен краткий обзор основных задач, решаемых при создании адаптивных суррогатных моделей (раздел 2), сформулирована новая задача снижения размерности данных при наличии "объясняющих переменных" (предикатов), возникающая, например, при построении суррогатных моделей для распределенных характеристик (раздел 3). В разделе 4 предложено решение этой задачи.

2. ОСНОВНЫЕ ПРОБЛЕМЫ СОЗДАНИЯ СУРРОГАТНЫХ АДАПТИВНЫХ МОДЕЛЕЙ

Основная концепция создания суррогатных адаптивных моделей заключается в следующих положениях [3]:

1) Характеристика объекта (Z), определяющая свойства объекта в некоторых условиях, может быть описана в виде функциональной зависимости $Z = F(X, Y)$, где переменная X описывает сам объект, а переменная Y задает условия функционирования объекта (параметры управления объектом, параметры внешней среды). Например, аэродинамические характеристики самолета (коэффициенты сил, моментов, сопротивлений и др.) в условиях крейсерского полета являются функцией, зависящей от формы поверхности самолета (X) и параметров режима полета и управления (Y) (например, скорости, углов атаки и скольжения и и др.).

2) Функция F является неизвестной, и для ее вычисления проводятся натурные или вычислительные эксперименты, то есть значения функции вычисляются с использованием моделей. Пусть M – некоторая модель (способ, функция), позволяющая вычислять приближенное значение $Z_M = F_M(X, Y)$ характеристики Z для входных данных (X, Y). Если функции F_M и F близки друг к другу в некоторой метрике:

$$F_M(X, Y) \approx F(X, Y), \tag{1}$$

то можно считать, что модель M достаточно адекватна реальности.

3) Имеется некоторое количество измерений

$$\Sigma = \{(X_i, Y_i, Z_i = F_i(X_i, Y_i)), i = 1, 2, \ldots\}, \tag{2}$$

где значение $Z_i = F_i(X_i, Y_i)$ характеристики Z получено методом M_i для объекта, имеющего описание X_i, в условиях функционирования Y_i. Предполагается, что имеющиеся измерения имеют приемлемую точность, то есть, $F_i(X_i, Y_i) \approx F(X_i, Y_i)$.

4) По известному множеству Σ (2) с использованием тех или иных математических методов анализа и обработки данных строится функция $F_S(X, Y)$, значение которой принимаются в качестве приближенного значения характеристики Z для объекта с описанием X в условиях функционирования Y.

Если все значения в множестве Σ (2) получены при помощи одной и той же модели M, и

$$F_S(X, Y) \approx F_M(X, Y), \tag{3}$$

то построенная функция F_S может рассматриваться как "заменитель" (суррогат) функции F_M. Методы вычисления характеристик с использованием таким образом построенных функций носят название "суррогатные модели".

Если получение данных с помощью модели M (функции F_M) является существенно более затратным (по времени, стоимости и/или другим показателям) по сравнению с построенной моделью S (функцией F_S), то построенную суррогатную модель $S = S(M)$ можно в дальнейшем использовать вместо модели M для вычисления приближенных значений неизвестной функции $F(X, Y)$.

Базируясь на вышеизложенной концепции, можно определить основные этапы и задачи построения суррогатных моделей.

Этап 1. Идентификация класса рассматриваемых объектов и создание математических Моделей описания объектов и условий их функционирования. Как и при построении любой модели M достаточно сложного объекта, определяющей функцию $F_M(X, Y)$, необходимо использовать некоторые модели для описания аргументов (X, Y) функции F: модели описания объектов рассматриваемого класса и модели условий их функционирования. При построении суррогатных моделей, основанных на математических методах анализа и обработки данных, необходимо иметь достаточно компактные описания входных данных, обеспечивая при этом достаточную адекватность. Например, детальное описание поверхности самолета, состоящее из десятков тысяч чисел, необходимо заменить небольшим числом геометрических характеристик объекта (порядка десятков и сотен), отражающих наиболее существенные (с точки зрения решаемой инженерной задачи) свойства объекта.

Построение "низкоразмерных" параметрических моделей для описания условий функционирования носит, как правило, предметно-ориентированный характер. Например, в задачах аэродинамического проектирования, в крейсерском режиме полета условия набегающего потока описываются несколькими параметрами (числа Маха и Рейнольдса, углы атаки и скольжения и др.), а для учета турбулентности могут использоваться низкорейнольдовские (k, ε) модели (или даже более простые алгебраические модели для пути смещения, как это сделано в широко используемом промышленном пакете вычислительной аэродинамики SRAR-CD) [7]. Но модели описания объектов могут строиться на универсальных когнитивных технологиях, основанных на анализе данных.

Этап 2. Создание консолидированных гармонизированных данных. Имеющиеся данные могли быть получены с помощью разных методов и моделей, для разных условий и с разной точностью. На основании таких данных могут быть построены так называемые консолидированные (гармонизированные) данные, в которых для каждого значения аргумента имеется ровно одно измерение, которое характеризуется единственным точностным параметром. Общая проблема получения консолидированных гармонизированных данных является особенно важной при построении суррогатных моделей, основанном на анализе и обработке данных.

Этот этап может включать в себя также планирование и проведение дополнительных вычислительных экспериментов для получения недостающих данных или повышения точности уже имеющихся данных. Результатом этого этапа является множество консолидированных данных

$$\Sigma_{cons} = \{(X, Y, Z = F_{cons}(X, Y)), (X, Y) \in D_{cons}\}, \tag{4}$$

где D_{cons} состоит из множества значений аргумента (X, Y), для которых имеются консолидированные данные, а F_{cons} обозначает результат построения консолидированных данных. Можно также считать, что имеется метод (способ) M_{cons} получения консолидированных данных.

Так же, как и для исходных данных, предполагается приемлемая точность консолидированных данных:

$$F_{cons}(X, Y) \approx F(X, Y), (X, Y) \in D_{cons}. \tag{5}$$

Этап 3. Создание суррогатной модели объекта. С учетом предположений (1), (5), можно рассматривать множество Σ_{cons} (4) как множество приближенных известных значений искомой неизвестной функции $F(X,Y)$. Поэтому задача построения суррогатной модели может рассматриваться как задача аппроксимации, то есть как задача построения аппроксимирующей функции

$$F_{appr}(X,Y) = F_{appr}(X,Y|\Sigma_{cons}), \tag{6}$$

которая приближенно вычисляет значение характеристики Z в заданной точке (X,Y) по множеству Σ_{cons} (4) приближенных известных значений функции $F(X,Y)$ в конечном числе точек $(X,Y) \in D_{cons}$. Построенная функция (6) и принимается в качестве суррогатной модели $F_S(X,Y)$.

Этап 4. Валидация и оценивание точности созданной суррогатной модели. На этом этапе проверяется адекватность созданной суррогатной модели F_S, то есть оценивается величина погрешности в соотношении (3) с использованием независимых высокоточных данных (High Fidelity Data), которые принимаются в качестве эталонных реальных данных. На этом этапе должна решаться также задача прогноза погрешности (3) для конкретных входных данных (X,Y).

Рассмотрим более подробно две математические задачи, решаемые при создании суррогатных моделей:

- задача снижения размерности данных при создании математических Моделей описания объектов,
- задача аппроксимации зависимостей при создании суррогатных моделей.

Задача снижения размерности при создании математических Моделей описания объектов [3–6] формулируется следующим образом. Пусть $B = \{b\}$ есть множество рассматриваемых объектов. Для каждого объекта $b \in$ B имеется его детальное описание $X = X(b)$ с максимальной степенью детальности. В реальных задачах размерность N вектора X может достигать тысяч чисел.

Зафиксируем некоторый набор параметров объекта $H_{mod}(b)$, определяющий отображение

$$H_{mod}(b) : B \to G_{mod}, \tag{7}$$

где множество

$$G_{mod} = \{H_{mod}(b), b \in B\}, \tag{8}$$

являющееся образом множества B при отображении H_{mod}, является фактор-пространством множества объектов B, определяемое отображением H_{mod}.

Очевидно, что в общем случае существует целое множество объектов:

$$B_{mod}(h) = \{b \in B : H_{mod}(b) = h\}$$

с одним и тем же набором параметров h, и отображение H_{mod} определяет разбиение пространства B на непересекающиеся подмножества $B_{mod}(h), h \in G_{mod}$.

Для каждого объекта $b \in B$ выберем (определим некоторым образом) **единственный** объект

$$b_{mod} = b_{mod}(b) \in B_{mod}(H_{mod}(b)),$$

называемый **модельным объектом**, соответствующим исходному объекту b, и обозначим

$$B_{mod} = \{b_{mod}(b), b \in B\} \subset B$$

множество всех модельных объектов. По построению, между множествами B_{mod} и G_{mod} существует взаимно-однозначное соответствие, определяемое прямым (7) и обратным отображениями:

$$H^{-1}_{\text{mod}} : G_{mod} \to B_{mod}, \tag{9}$$

с помощью которых модельный объект определяется как

$$b_{mod}(b) = H^{-1}_{\text{mod}}\,(H_{mod}(b)). \tag{10}$$

Модельный объект (10), построенный по объекту b с помощью пары отображений $\{H_{mod}, H^{-1}_{\text{mod}}\}$, будем называть также **модельным представлением** или **модельным аналогом** объекта b.

По построению, множество модельных объектов B_{mod} является многообразием в пространстве объектов B, параметризованным с помощью отображения H^{-1}_{mod}, определенного на факторпространстве G_{mod} (8). Обратное отображение H^{-1}_{mod} (9) определяет также "алгоритм восстановления", позволяющий для каждого исходного объекта $b \in B$ строить детальное описание

$$X_{mod} = X(b_{mod}(b)) = X(H^{-1}_{\text{mod}}\,(H_{mod}(b))) \tag{11}$$

соответствующего модельного объекта $b_{mod}(b)$ (10).

Если вектор $H_{mod}(b)$ имеет небольшую (по сравнению с размерностью N детального описания $X(b)$) размерность n, и если все объекты b $\in B_{mod}(h)$, имеющие одно и то же значение вектора параметров модели h, имеют близкие детальные описания, то есть если

$$X(b') \approx X(H^{-1}_{\text{mod}}\,(h)),$$

для всех $h \in G_{mod}$ и $b' \in B_{mod}(h)$, то пара отображений $\{H_{mod}, H^{-1}_{\text{mod}}\}$ определяет процедуру снижения размерности (сжатия) описания объекта:

• отображение H_{mod} определяет процедуру снижения размерности (сжатия) детального описания объекта:

$$X(b) \to \quad H_{mod}(b),$$

и величину $H_{mod}(b)$ можно считать "сжатым" описанием объекта;

• отображение H^{-1}_{mod} определяет процедуру восстановление детального описания объекта по его сжатому описанию:

$$H_{mod}(b) \to X(H^{-1}_{\text{mod}}\,(H_{mod}(b))),$$

и погрешность приближенного равенства

$$X(b) \approx X(H^{-1}_{\text{mod}}\,(H_{mod}(b))) \tag{12}$$

определяет точность процедуры сжатия.

Однако постановка **задачи снижения размерности**, решаемой при создании суррогатных моделей, имеет ряд особенностей:

• к стандартным требованиям близости (12) исходного описания и описания, восстановленного после сжатия, могут добавляться различные требования, например, требования "функциональной" близости этих описаний:

$$F(X(b), Y) \approx F(X(H^{-1}_{\text{mod}}\,(H_{mod}(b))), Y), \tag{13}$$

• класс рассматриваемых объектов $B = \{b\}$ не имеет, как правило, точного описания и определяется конечным множеством его "представителей" (прототипов), задаваемых множеством их детальных описаний

$$\mathbf{X} = \mathbf{X}_T = \{X(b_t), t = 1, 2, \ldots, T\}.$$

В работах [8, 9], см. также [10–16], приведен краткий обзор основных используемых подходов и методов для решения задачи снижения размерности и предложен общий подход к построению процедур снижения размерности данных, описывающих объекты.

Задача аппроксимации зависимостей при создании суррогатных моделей [3, 4] состоит в построении функции $F_{appr}(X, Y)$ (6), которая может быть принята в качестве приближенного значения неизвестного истинного значения $F(X, Y)$ характеристики Z (см. (3)).

Технология построения суррогатных моделей основана на использовании цепочки преобразований.

Преобразование 1. С использованием Модели описания объектов, вместо исходного объекта b, имеющего детальное описание $X = X(b)$, рассматривается модельный объект $b_{mod}(b)$, имеющий описание $X_{mod} = X(b_{mod}(b))$ (11). Свойства модели обеспечивают приближенное равенство (13), которое позволяет свести задачу вычисления значения характеристики $F(X(b), Y)$ объекта b к задаче оценки характеристики модельного объекта $F(X(b_{mod}), Y)$.

Так как $X(b_{mod})$ зависит только от модельного объекта b_{mod} только через вектор параметров модели $h = H_{mod}(b)$, то, с учетом обозначения

$$F_{mod}(h, Y) = F(X(H^{-1}_{\text{mod}}(h)), Y), \tag{14}$$

соотношение (13) может быть записано с помощью следующего соотношения:

$$F(X(b), Y) \approx F_{mod}(h, Y), h = h(b) = H_{mod}(b). \tag{15}$$

Тем самым, задача построения модели для вычисления $F(X(b), Y)$ может быть заменена на задачу построения модели для вычисления функции $F_{mod}(h(b), Y)$ (14), зависящей от аргумента h с существенно более низкой, по сравнению с X, размерностью.

Преобразование 2. Пусть M – некоторая существующая модель, позволяющая вычислять приближенное значение $F_M(X(b), Y)$ характеристики Z. Из условия (1) следует, что для модельных объектов имеет место приближенное равенство:

$$F_M(X(b_{mod}(b)), Y) \approx F_{mod}(h, Y), h = h(b) = H_{mod}(b).$$

Обозначив

$$F_{Mmod}(h, Y) = F_M(X(H^{-1}_{\text{mod}}(h)), Y),$$

получаем:

$$F_{Mmod}(h, Y) \approx F_{mod}(h, Y), h = h(b) = H_{mod}(b). \tag{16}$$

Тем самым, модель для вычисления функции $F_{mod}(h(b), Y)$ может быть заменена моделью для вычисления функции $F_{Mmod}(h(b), Y)$.

Пусть имеется множество данных Σ результатов экспериментов по вычислению характеристики Z с использованием различных моделей для множества объектов B_{cons}, по которым построено множество консолидированных данных Σ_{cons} (4), состоящих из множества значений характеристики Z, вычисленных для множества значений аргумента $(X, Y), (X, Y) \in D_{cons}$.

Рассматривая в качестве модели M метод получения консолидированных данных M_{cons} (точность которого заведомо не ниже точности каждого частного источника данных), приближенное равенство (16) можно записать в виде:

$$F_{cons}(X(b), Y) \approx F_{mod}(H_{mod}(b), Y), (X(b), Y) \in D_{cons}.$$

Обозначив

$$F_{cons-mod}(h, Y) = F_{cons}(X(H^{-1}_{\text{mod}}(h)), Y), \tag{17}$$

$$D_{cons-mod} = \{(H_{mod}(b), Y) : (X(b), Y) \in D_{cons}\} \tag{18}$$

получаем:

$$F_{cons-mod}(h, Y) \approx F_{mod}(h, Y), (h, Y) \in D_{cons-mod}, \tag{19}$$

и, следовательно, имеется множество приближенных значений $F_{cons-mod}(h, Y)$ (17) неизвестной функции $F_{mod}(h, Y)$ для множества значений аргументов $(h, Y) \in D_{cons-mod}$ (18).

Преобразование 3. Пусть по множеству известных приближенных значений (18) построена функция $F_{appr}(h, Y)$, достаточно точно аппроксимирующая неизвестную функцию $F_{cons-mod}(h, Y)$ (19) на множестве значений аргументов $D_{mod} = \{(h, Y) : Y \in G_{mod}\}$:

$$F_{appr}(h, Y) \approx F_{cons-mod}(h, Y), (h, Y) \in D_{mod}. \tag{20}$$

В результате цепочки преобразований и построений, обеспечивающих соотношения (15), (16), (19) и (20), может быть построена суррогатная модель M_{surr}, вычисляющая приближенное значение $F(X, Y)$ характеристики Z с помощью функции:

$$F_{surr}(X(b), Y) = F_{appr}(H_{mod}(b), Y).$$

В итоге, исходная задача построения суррогатной модели для приближенного вычисления функции $F(X(b), Y)$ может быть сведена к **Задаче построения аппроксимирующей функции** $F_{appr}(h, Y)$ для вычисления значения $F_{mod}(h, Y)$.

Типичным образом, задачи аппроксимации, возникающие при создании суррогатных моделей, имеют ряд особенностей, описанных в [3, 4], поэтому при создании суррогатных моделей приходится разрабатывать новые комбинированные методы аппроксимации, сочетающие в себе методы искусственного интеллекта (Искусственных Нейронных Сетей [17]) и традиционные математические методы аппроксимации и анализа данных [18, 19].

Среди характеристик, описывающих объект, могут присутствовать характеристики $Z = F(X, Y)$, в которых Z является непрерывной функцией некоторого аргумента l. Например, одной из аэродинамических характеристик самолета является распределенная нагрузка по крылу (Wing span load distribution) при крейсерском режиме полета, определяемая как функция $Z(l)$, определенная вдоль размаха крыла. При построении суррогатных моделей для таких характеристик возникают новые постановки задач аппроксимации и снижения размерности, рассмотренные в следующем разделе.

3. ЗАДАЧА АППРОКСИМАЦИИ НЕПРЕРЫВНЫХ ЗАВИСИМОСТЕЙ

Пусть имеется множество консолидированных данных Σ_{cons} (4), состоящих из значений характеристики Z, вычисленных для множества значений аргумента $(X, Y), (X, Y) \in D_{cons}$. Мы рассматриваем здесь случай, когда переменная Z является функцией одномерного аргумента

l, областью значений которого без ограничения общности можно считать отрезок [0, 1]. Тем самым, соотношение (2) принимает вид

$$\Sigma = \{(X_i, Y_i, Z_i(l) = F_i(X_i, Y_i, l)), i = 1, 2, \ldots, T\},$$

где $l \in [0, 1]$. Если функции $Z(l)$ являются результатами каких-либо натурных или вычислительных экспериментов, то они представляются в виде набора значений этих функций для некоторого набора аргументов $\{l_1 = 0, l_2, \ldots, l_{m-1}, l_m = 1\}$, и равенство (2) может быть записано в виде:

$$\Sigma = \{(X_i, Y_i, Z_{ik} = F_i(X_i, Y_i, l_k)), k = 1, 2, \ldots, m; i = 1, 2, \ldots, T\}.$$

Непосредственное применение подхода, описанного в предыдущем разделе, приводит к необходимости построения аппроксимационных зависимостей $F_{kappr}(h, Y)$ для каждой k-ой компоненты $Z(l_k) = F(X(H^{-1}_{\text{mod}}(h)), Y, l_k), k = 1, 2, \ldots, m$, вектора

$$\mathbf{Z} = (Z(l_1), Z(l_2), \ldots, Z(l_m)), \tag{21}$$

состоящего из значений функции $Z(l)$, на основании множества значений

$$\Sigma_k = \{(X_i, Y_i, Z_{ik} = F_i(X_i, Y_i, l_k)), i = 1, 2, \ldots, T\}.$$

Пусть $\Lambda \subset R^m$ есть множество m-мерных векторов, отвечающих различным объектам $b \in B$ и различным условиям Y функционирования объекта. В большинстве реальных приложений множество Λ или само имеет размерность, меньшую m, либо может быть аппроксимировано многообразием меньшей размерности. Однако при построении независимых аппроксимаций для каждой k-ой компоненты $Z(l_k)$ вектора $\mathbf{Z}$ (21) невысокая собственная размерность множества векторов Λ не учитывается, что приводит к существенным ошибкам аппроксимации характеристики Z, являющейся непрерывной функцией. Поэтому предлагается иной подход к построению аппроксимаций, лишенный указанного недостатка и основанный на параметризации многообразия Λ.

В статье [9] подробно рассмотрен вопрос о построении такого многообразия, основанного на процедурах снижения размерности описания векторов $\mathbf{Z} \in \Lambda$ с использованием множества данных

$$\mathbf{Z}_T = \{\mathbf{F}_i(X_i, Y_i) = (F_i(X_i, Y_i, l_1), F_i(X_i, Y_i, l_2), \ldots, F_i(X_i, Y_i, l_m)), i = 1, 2, \ldots, T\}. \tag{22}$$

Пусть Π – процедура снижения размерности, преобразующая m-мерные векторы $\mathbf{z} \in \Lambda$ в s-мерные векторы $\mathbf{u} = \Pi(\mathbf{z}), s < m$, и Π^- – процедура "восстановления", позволяющая строить по исходному "сжатому" s-мерному вектору $\mathbf{u}$ m-мерный восстановленный вектор $\Pi^-(\mathbf{u})$. В совокупности эти процедуры обеспечивают близость исходного и восстановленного векторов:

$$\mathbf{z} \approx \Pi^-(\Pi(\mathbf{z})). \tag{23}$$

Параметрическое многообразие

$$\Lambda_{appr} = \{\Pi^-(\mathbf{u}) : \mathbf{u} \in \mathbf{U} \subset R^s\}, \quad \mathbf{U} = \{\Pi(\mathbf{z}) : \mathbf{z} \in \Lambda \subset R^m\},$$

имеет размерность $s < m$ и, в соответствии с (23), аппроксимирует многообразие Λ.

Процедура аппроксимации, результат которой лежит (приближенно) на многообразии Λ, состоит из следующих шагов:

Шаг 1. Строится множество

$$\mathbf{U}_T = \{\ \mathbf{u}_i = \Pi\mathbf{z}_i = \Pi\mathbf{F}_i(X_i, Y_i), i = 1, 2, \ldots, T\},$$

состоящее из результатов применения процедуры снижения размерности Π к множеству $\mathbf{Z}_T$ (22).

Шаг 2. По множеству $\mathbf{U}_T$ (23) строится аппроксимация $\Phi_{appr}(X, Y)$ для многомерной s-мерной функции

$$\Phi(X, Y) = (\Phi_1(X, Y), \Phi_2(X, Y), \ldots, \Phi_s(X, Y)) = \Pi\mathbf{F}(X, Y).$$

Шаг 3. Многомерная m-мерная функция

$$\mathbf{F}_{appr}(X, Y) = \Pi^{-}\Phi_{appr}(X, Y)$$

выбирается в качестве аппроксимации для исходной функции $\mathbf{F}(X, Y)$.

Однако у рассмотренной вспомогательной задачи снижения размерности есть особенность, которая отличает ее от стандартных задач снижения размерности. А именно, наряду с компонентами $\mathbf{F}(X, Y)$ множества $\mathbf{Z}_T$ (22), использованного при построении аппроксимирующего многообразия, нам известны значения аргументов (X, Y) этих функций. Эти величины могут рассматриваться как сопутствующие переменные (предикаты), "объясняющие" значения элементов множества $\mathbf{Z}_T$ (22). Задача снижения размерности при наличии предикатов рассмотрена в следующем разделе.

4. СНИЖЕНИЕ РАЗМЕРНОСТИ ПРИ НАЛИЧИИ ПРЕДИКАТОВ

Исходя из сути процедур снижения размерности, интуитивно ясно, что чем меньше изменчивость векторов в множестве $\mathbf{Z}_T$ (22), используемом при построении процедур снижения размерности, тем лучшего качества процедур можно достичь.

Рассмотрим задачу снижения размерности множества m-мерных векторов

$$\mathbf{Z}_T = \{\mathbf{z}_i \in R^m, i = 1, 2, \ldots, T\}, \tag{24}$$

в котором значение каждого m-мерного вектора $\mathbf{z}_i$ получено в условиях, которые характеризуются значением некоторого параметра $\theta_i \in \Theta, i = 1, 2, \ldots, T$. Предлагается подход к построению процедур снижения размерности, позволяющий учесть имеющуюся дополнительную информацию о векторах $\mathbf{z} \in \mathbf{Z}_T$ (значения объясняющих переменных θ) для снижения изменчивости множества векторов $\mathbf{Z}_T$.

Предложенный подход состоит из нескольких шагов:

Шаг 1. По множеству пар $\{(\mathbf{z}_i, \theta_i), i = 1, 2, \ldots, T\}$ строится регрессия векторов $\mathbf{z}$ на векторы θ. В предположении однородной вероятностной структуры этого множества пар, наилучшей регрессией (при квадратичной функции потерь) является условное математическое ожидание $\mathbf{Z}(\theta) = E(\mathbf{z}|\theta)$. Без предположения о такой структуре данных, можно рассмотреть различные процедуры, используемые в анализе данных [20], для построения зависимостей $\mathbf{z}_T(\theta)$ (например, процедуру построения линейной регрессии векторов $\mathbf{z}$ на θ).

Шаг 2. Вычисляются остатки регрессии:

$$\mathbf{z}_{iT} = \mathbf{z}_i - \mathbf{z}_T(\theta_i), i = 1, 2, \ldots, T. \tag{25}$$

Шаг 3. К выборке, состоящей из остатков регрессии (25), применяется процедура снижения размерности, состоящая из пары процедур сжатия/восстановления (Π, Π^{-}). Пусть векторы

$$\mathbf{u}_{iT} = \Pi\mathbf{z}_{iT} = \Pi(\mathbf{z}_i - \mathbf{z}_T(\theta_i)), i = 1, 2, \ldots, T,$$

принимаются в качестве результата процедуры сжатия векторов (25).

Шаг 4. Векторы

$$\mathbf{v}_{iT} = \Pi^{-}\mathbf{u}_{iT} + \mathbf{z}_T(\theta_i) = \Pi^{-}\Pi(\mathbf{z}_i - \mathbf{z}_T(\theta_i)) + \mathbf{z}_T(\theta_i), i = 1, 2, \ldots, T,$$

принимается в качестве результата процедуры восстановления для векторов $\mathbf{z} \in \mathbf{Z}_T$.

Приведем в качестве иллюстрации предложенного подхода результаты вычислительных экспериментов с реальными данными, связанными с описанием распределенных аэродинамических характеристик. Рассматривалась реальная выборка (24), состоящая из $T = 12251$ m $= 21$-мерных векторов, описывающих распределенную нагрузку по крылу (Wing span load distribution) самолета при крейсерском режиме полета в заданных $m = 21$ фиксированных точках, распределенных равномерно по размаху крыла. Эти данные были получены с помощью CFD-кода для различных компоновок и различных значений параметров режима полета (чисел Маха и углов атаки), описываемых в совокупности векторами $\{\theta_i, i = 1, 2, \ldots, T\}$.

Сравнивалось две процедуры:

• процедура снижения размерности (Π, Π^{-}) без учета предикатов (названная для краткости стандартной),

• предложенная процедура снижения размерности (также с использованием пары (Π, Π^{-})), но с учетом предикатов (названная для краткости усовершенствованной).

Рассматривалось 4 метрики, определяющие расстояния между исходными и восстановленными векторами в R^{21}: максимальная ошибка (l_∞), максимальная относительная ошибка (lr_∞), средняя ошибка (l_1) и средняя относительная ошибка (lr_1). В таблице приведены значения соответствующих ошибок, усредненные по элементам выборки, иллюстрирующие эффективность предложенных процедур.

Метрики	**Процедура**	
	Стандартная	**Усовершенствованная**
Максимальная ошибка	0.0340	0.0303
Максимальная относительная ошибка	0.1543	0.0734
Средняя ошибка	0.0038	0.0031
Средняя относительная ошибка	0.0076	0.0061

Предложенные процедуры были использованы в быстрых аэродинамических моделях самолета, построенных на основе когнитивной технологии быстрых расчетов [21–24].

СПИСОК ЛИТЕРАТУРЫ

1. Lucia, D.J. Beran, P.S. and Silva, W.A., Reduced-order Modeling: New Approaches for Computational Physics, *Progress in Aerospace Sciences*, 2004, vol. 40, no. 1-2, pp. 51–117.

2. Antoulas, A.C., Sorensen, D.C., and Gugercin, S., A Survey of Model Reduction Methods for Large-scale Systems, *Structured Matrices in Operator Theory, Numerical Analysis, Control, Signal and Image Processing,' Contemporary Mathematics*, AMS publications, 2001.

3. Кулешов, А.П., Когнитивные технологии в основанных на данных адаптивных моделях сложных объектов, *Информационные технологии и вычислительные системы*, 2008, вып. 1, с. 95–106.

4. Кулешов, А.П., Технология быстрого вычисления характеристик сложных технических объектов, *Информационные технологии. Прил.*, 2006, № 3, стр. 4–11.

5. Кулешов, А.П., Основные принципы технологии быстрых расчетов, основанной на знаниях, *Тезисы докладов Международной научно-технической конференции "Информационные технологии и математическое моделирование в науке, технике и образовании"*, Сицилия (Италия), 2006.

6. Кулешов, А.П., Информационные технологии в проблеме создания моделей сложных объектов, *Вторая Международная конференция "Системный анализ и информационные технологии" САИТ-2007*, (10–14 сентября 2007 г., г. Обнинск, Россия), Труды конференции, 2007, т. 1, стр. 14–16.

7. Вышинский, В.В., Судаков, Г.Г., Применение численных методов в задачах аэродинамического проектирования, *Труды ЦАГИ*, 2007, вып. 2673, стр. 1–142.

8. Бернштейн, А.В., Кулешов, А.П., Задачи снижения размерности моделей сложных объектов, *Вторая Международная конференция "Системный анализ и информационные технологии" САИТ-2007*, (10–14 сентября 2007 г., г. Обнинск, Россия), Труды конференции, 2007, т. 1, стр. 243–247.

9. Бернштейн, А.В., Кулешов, А.П., Когнитивные технологии в проблеме снижения размерности описания геометрических объектов, *Информационные технологии и вычислительные системы*, 2008, вып. 2.

10. Jackson, J.E., *A User's Guide to Principal Components, Wiley Series in Probability and Mathematical Statistics*, New York: Wiley, 1991.

11. DeMers, D. and Cottrell, G.W., Nonlinear Dimensionality Reduction, in *Advances in Neural Information Processing Systems*, Hanson, S.J., Cowan, J.D., and Giles, C.L., Eds., San Mateo: Morgan Kaufman, 1993, vol. 5, pp. 580–587.

12. Huber, P.J., Projection Pursuit, *Annals of Statistics*, 1985, 13(2), pp. 435–475.

13. Fodor, I.K., A Survey of Dimension Reduction Technique, *LLNL Technical Report*, June 2002, UCRL-ID-148494.

14. Hastie, T.J. and Stuetzle, W., Principal Curves, *J. Am. Stat. Assoc.*, 1898, 84(406), pp. 502–516.

15. Айвазян, С.А., Бухштабер, В.М., Енюков, И.С., Мешалкин, Л.Д., *Прикладная статистика: классификация и снижение размерности*, М.: Финансы и статистика, 1989.

16. Vapnik, V., *Statistical Learning Theory*, New-York: Wiley, 1998.

17. Hornick, K., Stinchcombe, M., and White, H., Multilayer Feedforward Networks are Universal Approximators, *Neural Networks*, 1989, vol. 2, no. 5, pp. 359–366.

18. Коровкин, П.П., *Линейные операторы и теория приближений*, М., 1959.

19. Ахиезер, Н.И., *Лекции по теории аппроксимации*, М., 1965.

20. Айвазян, С.А., Енюков, И.С., Мешалкина, Л.Д., *Прикладная статистика. Исследование зависимостей*, М.: Финансы и статистика, 1985.

21. Bernstein, A.V., Kuleshov, A.P., Sviridenko, Y.N., and Vyshinsky, V.V., Fast Aerodynamic Model for Design Technology, *Proceedings of West-East High Speed Flow Field Conference*, November 19–22, 2007, Moscow, Russia, http://wehsff.imamod.ru/pages/s7.htm.

22. Bernstein, A.V., Kuleshov, A.P., Sviridenko, Y.N., and Vyshinsky, V.V., Fast Aerodynamic Model for Design Technology, *Workbook "West-East High Speed Flow Field Conference"*, November 19–22, 2007, Moscow, Russia, pp. 125–126.

23. Бернштейн, А.В., Вышинский, В.В., Кулешов, А.П., Свириденко, Ю.Н., Применение искусственных нейронных сетей для определения нагрузок по крылу пассажирского самолета на режиме крейсерского полета, *Труды Центрального аэрогидродинамического института им. проф. Н.Е. Жуковского*, Выпуск № 2678 *"Применение искусственных нейронных сетей в задачах прикладной аэродинамики"*, Москва, 2008.

24. Бернштейн, А.В., Вышинский, В.В., Кулешов, А.П., Свириденко, Ю.Н., Быстрый метод аэродинамического расчета для задач проектирования, *Труды Центрального аэрогидродинамического института им. проф. Н.Е. Жуковского*, Выпуск № 2678 *"Применение искусственных нейронных сетей в задачах прикладной аэродинамики"*, Москва, 2008.