Интеллектуальный анализ данных в теории надежности

Бернштейн А.В.
Лаборатория когнитивных методов анализа данных и моделирования
Институт Системного Анализа РАН
Пр-т 60-летия Октября, 9, г. Москва
Россия
a.bernstein@irias.ru

Аннотация

Доклад посвящен обзору задач интеллектуального анализа данных, возникающих при построении математических моделей объектов и процессов в теории надежности, основанных на данных (суррогатных моделей). Особое внимание будет уделено использованию когнитивных методов, объединяющих совокупность методов, алгоритмов и программ анализа данных, моделирующих познавательные способности человеческого мозга, для построения суррогатных моделей. Будут рассмотрены также проблемы использования суррогатных моделей в компьютерных системах проектирования и поддержки принятия решений в условиях нечетких входных данных и взаимосвязей.

1 Математические модели теории надежности в системах проектирования

Математические методы в теории надежности используются для построения и анализа аналитических моделей, описывающих различные процессы, происходящие с исследуемыми объектами, при их функционировании во времени и/или под воздействием окружающей среды и влияющие на их надежность. Построенные аналитические модели позволяют предсказывать те или иные характеристики объектов (времена жизни в определенных условиях, запас прочности при воздействии на объект внешних воздействий, и т.п.) в зависимости от характеристик объекта и внешних условий.
В процессе проектирования сравниваются различные технические решения, касающиеся структуры и параметров объекта, механизмов его функционирования и других элементов объекта, и надежность объекта является одним из важнейших факторов, учитываемых при сравнении технических решений. Компьютерные системы проектирования создаются для сокращения времени проектирования и числа дорогостоящих натурных экспериментов и являются, по существу, системами моделирования, в состав которых входят и модели, позволяющие вычислять надежностные характеристики проектируемого объекта.
Традиционно в моделировании используются математические модели, основанные на «физике процессов» и описывающие физические процессы и явления, происходящие при функционировании объекта, сложными дифференциальными уравнениями в частных производных с граничными условиями, решаемые численными методами, имеющими значительную вычислительную трудоемкость, как самих расчетов, так и подготовки исходных данных и расчетных сеток. Это существенно сокращает возможности использования моделей, основанных на «физике процессов», особенно на стадии предварительного (концептуального) проектирования, на которой рассматривается очень большое количество вариантов решений и высока цена неправильно выбранного решения.

2 Модели, основанные на данных (суррогатные модели)

В последние годы стали развиваться математические модели, основанные на данных - результатах натурных и/или вычислительных экспериментов, проведенных с различными объектами рассматриваемого класса. Другими словами, модели «обучаются» по множеству прототипов входных и выходных данных и фактически имитируют (заменяют) как источники получения данных, основанные на некоторой исходной модели, так и сами модели, созданные на основе изучения физики процессов. Поэтому, такие адаптивные модели иногда называют также метамоделями или суррогатными моделями. Пример построения суррогатных моделей для анализа прочности конструкций содержится в докладе [5].

Технология построения суррогатных моделей и используемые методы [1], [6], [7], [9] основаны на синергии методов предметной области и когнитивных технологий, базирующихся на достижениях общенаучных дисциплин (математики, искусственного интеллекта и анализа данных, информационных технологий).

3 Интеллектуальный анализ данных при построении суррогатных моделей

При проектировании объектов решаются две основные задачи:

- анализ конкретного варианта построения объекта, то есть, вычисление различных характеристик объекта по заданным цифровому описанию объекта, параметрам управления объектом и параметрам внешней среды (например, среды функционирования),
- оптимизация структуры объекта, то есть построение цифрового описания объекта с требуемыми (наилучшими) свойствами при наличии ограничений. С точки зрения задачи оптимизации, задача анализа состоит в построении функции отклика. Задача оптимизации включает в себя также задачу целенаправленной автоматической генерации вариантов цифровых описаний объектов,

для решения которых необходимо строить и оптимизировать аналитические модели объектов. Сложности построения и оптимизации моделей объектов, построенных по данным, обусловлены, прежде всего, высокой размерностью цифровых описаний объектов (включающих в себя 3D-описания геометрии объектов, свойства материалов и т.п.). Высокая размерность описания объектов (достигающая тысяч чисел) существенно затрудняет построение по данным функций отклика, зависящих от векторов высокой размерности, и оптимизацию в пространстве таких векторов. Как правило, цифровые описания объектов рассматриваемого класса лежат вблизи многообразий существенно меньшей размерности, и необходимо «оставаться» вблизи этих многообразий при генерации новых объектов (в частности, в процессе оптимизации). Перечислим основные математические задачи анализа и обработки данных, используемые в процессе построения и оптимизации суррогатных моделей [1], [7]:

- определение внутренней размерности множества данных и построение процедур снижения размерности (построение аппроксимирующих многообразий меньшей размерности);
- построение многомерных нелинейных аппроксимирующих зависимостей;
- кластеризация и классификация данных;
- предсказание значений ошибок процедур;
- генерация (имитационное моделирование) многомерных данных, лежащих вблизи нелинейных многообразий меньшей размерности, и др.

Точные постановки этих задач, возникающие при построении и оптимизации суррогатных моделей, имеют ряд существенных особенностей по сравнению с классическими постановками. Эти особенности связаны как со спецификой предметных областей, так и с необходимостью взаимосвязанного решения различных задач, когда выходные данные одной частной задачи являются входными данными для другой задачи, и целевые функции для частных задач нельзя определить независимо. Перечислим кратко особенности постановок некоторых задач анализа данных, возникающих при построении и оптимизации суррогатных моделей.

3.1 Задачи снижения размерности

Результат решения задачи снижения размерности используется как начальный этап для других задач при построении и оптимизации суррогатных моделей. Процедуры снижения размерности входных или выходных данных должно обеспечивать близость (в различных метриках) не только между данными и их восстановленными (в результате последовательного применения процедур сжатия и восстановления) значениями, но и между значениями различных функционалов от них

[2]. Использование процедур снижения размерности для построения многообразия меньшей размерности, аппроксимирующего носитель данных, на котором и будет осуществляться оптимизация суррогатной модели, налагает жесткие условия на выбор размерности этого многообразия: при занижении размерности оптимизация будет проводиться лишь в узкой области значений параметров, а при завышении полученные оптимальные значения параметров могут не иметь содержательного смысла.

3.2 Задачи построения многомерных нелинейных аппроксимирующих зависимостей

Типичной является ситуация, когда входные (X) и выходные (Y) переменные в задаче аппроксимации (построения зависимости $Y = F(X)$) являются многомерными, и их носители лежат (приближенно) на многообразиях меньшей размерности. Формальное независимое снижение размерностей переменных X и Y может привести к плохому качеству построенной аппроксимации, так как не учитывается влияние ошибок снижения размерности переменной X на зависимую переменную Y, причем характер этого влияния (в силу неизвестности F) неизвестен. Поэтому задачу аппроксимации следует рассматривать как задачу восстановления значения $Y^*(X)$ пропущенной переменной Y в паре (X, Y) таким образом, чтобы, чтобы пара $(X, Y^*(X))$ лежала как можно ближе к многообразию , аппроксимирующему множество данных (X_i, Y_i). Такая постановка является симметричной по отношению к зависимой и независимой переменным, и позволяет решать обратную задачу построения зависимости $X(Y)$ (если зависимость между переменными является взаимно однозначной) или, в общем случае, задачу построения непараметрической регрессии X на Y.

3.3 Задачи кластеризации и классификации данных

Требования по точности построенной аппроксимирующей зависимости могут быть разными для различных областей изменения переменной X, причем эти области могут определяться в неявном виде, через значения неизвестной функции $F(X)$ или некоторых функционалов от нее. Учет этих требований приводит к необходимости построения нескольких аппроксимирующих зависимостей (например, зависимостей для некоторых кластеров данных), и окончательная процедура аппроксимации должна включать в себя классификатор, определяющий, какой именно частный аппроксиматор (или несколько аппроксиматоров, с последующей комбинацией их значений) необходимо выбрать для заданного значения входной переменной X. Так как процедуры кластеризации и классификации данных являются промежуточными в задаче построения аппроксимации, то к их решению предъявляются требования, существенно отличающиеся от требований в классических постановках этих задач.

4 Автоматизация процедур интеллектуального анализа данных

Модели, основанные на данных (прототипах), по своей сути могут гарантированно использоваться только для входных данных, которые подобны данным (прототипам), 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). Так как для большинства статистических задач не существует эффективных универсальных процедур для их решения, то суррогатные модели создаются специалистом в предметной области в анализе данных, знакомым с различными математическими методами и понимающим, как влияет структура данных на качество той или иной процедуры снижения размерности. 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 автоматического (автоматизированного) создания статистических процедур (генераторов 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 строятся на основе когнитивных технологий работы с данными и в процессе своей работы имитируют деятельность математика, проводя целенаправленно вычислительные эксперименты с различными встроенными процедурами анализа данных и синтезируя наиболее эффективную процедуру работы с заданным множеством данных.
Автоматические генераторы процедур построены для многих базовых процедур анализа данных (снижения размерности, аппроксимации и др.), и для своего создания потребовали развития новых математических процедур. Например, автоматический генератор процедуры снижения размерности основан на новых математических результатах [2], [3], [4] и позволяет для требуемой точности построить программно реализованную (в виде динамической библиотеки) процедуру снижения размерности с минимально возможной размерностью сжатых данных. В этой процедуре автоматически синтезирован ряд частных процедур (главных компонент, нейронных сетей, оптимизации и др.).

Список литературы

[1] Бернштейн, А.В., Кулешов, А.П. (2008). Математические методы в когнитивных инженерных технологиях. Обозрение прикладной и промышленной математики, сер. *Вероятность и статистика*. 15(3), 451 – 452.

[2] Бернштейн, А.В., Кулешов, А.П. (2008). Когнитивные технологии в проблеме снижения размерности описания геометрических объектов. *Информационные технологии и вычислительные системы*. 2, 6 – 19.

[3] Бернштейн, А.В., Бурнаев, Е.В., Дорофеев, Е.А., Свириденко, Ю.Н., Чернова, С.С. (2008). Каскадные процедуры снижения размерности. *Труды XI национальной конференции по искусственному интеллекту с международным участием* (Дубна, 2008), 1, 241 – 250.

[4] Бернштейн, А.В., Кулешов, А.П. (2008). Построение ортогональных нелинейных многообразий в задачах снижения размерности. *Труды VII Международной школы-семинара «Многомерный статистический анализ и эконометрика»* Цахкадзор, 2008, 25 – 27.

[5] Burnaev, E., Grihon, S. (2009). Construction of the Metamodeles in Support of Stiffened Panel Optimization. In the present book.

[6] Forrester, A.I.J., Sobester, A., Keane, A.J. (2008) Engineering Design via Surrogate Modelling. A Practical Guide. New-York: Wiley.

[7] Кулешов, А.П. (2008). Когнитивные технологии в адаптивных моделях сложных объектов. *Информационные технологии и вычислительные системы*, 1, 18 – 29.

[8] Кулешов, А.П. (2008). Задачи многомерного статистического анализа в системах компьютерного проектирования. *Труды VII Международной школы-семинара «Многомерный статистический анализ и эконометрика»* (Армения, Цахкадзор 2008), 60 – 61.

[9] Wang, G., Gary, Shan, S. (2007). Review of Metamodeling Techniques in Support of Engineering Design Optimization. J. Mech. Des. 129(3), 370-381.