Автоматическая генерация процедур интеллектуальной обработки данных

Чернова С. С.
Институт системного анализа РАН
пр-т 60-летия октября, 9, г. Москва, Россия
chernova@cpt-ran.ru

Маркевич С. В.
Институт проблем передачи информации РАН
Большой Каретный переулок, 19, г. Москва, Россия *markevich@irias.rchernova@cpt-ran.ru*

Аннотация

В докладе описывается способ построения автоматических генераторов процедур, которые создают программные модули, реализующие базовые процедуры анализа и обработки данных, существенно зависят от данных и могут быть использованы как внешние библиотеки при разработке специализированных инженерных приложений.

1 Введение

Повышение надежности функционирования технического объекта всегда было и остается актуальной, но трудно разрешимой проблемой. Многие базовые решения относительно элементной базы, структуры, материалов и других параметров принимаются на этапах раннего концептуального проектирования. Для принятия таких инженерных решений проводят расчет и анализ показателей, характеризующих поведение проектируемого объекта при различных условиях эксплуатации для различных вариантов его построения.

В настоящее время математическое моделирование, вычислительные эксперименты для исследования аналитических моделей создаваемого объекта и окружающей его среды стали одним из самых распространенных методов анализа и оптимизации структуры технических объектов. При расчете характеристик используют ранее накопленные данные о надежности элементов объекта, о надежности объектов-аналогов, о свойствах материалов, а также и другую информацию, имеющуюся к моменту расчета. Вычисленные характеристики в последствии могут добавляться в существующие базы данных и использоваться при последующих вычислительных экспериментах.

Развитие интеллектуальных средств анализа данных влекут за собой потребность в создании прикладных инженерных инструментальных средств, предназначенных для решения конкретных инженерных задач. Эти средства должны комбинировать современные методы обработки данных, базирующиеся на достижениях общенаучных дисциплин, и методы конкретной предметной области. Такие инструментальные средства должны «уметь» адаптироваться к специфике предметной области путем подбора процедур и алгоритмов обработки данных, которые являются наиболее эффективными для конкретного набора данных или проблемно-ориентированной постановки задачи.

Для сокращения времени проектирования и проведения массовых расчетов современные когнитивные технологии предлагают математические модели, основанные на данных. Такие модели строятся на основе данных - результатов натурных и/или вычислительных экспериментов, проведенных с различными объектами рассматриваемого класса, с минимальным привлечением знаний из предметной области (физики процессов). Другими словами, модели «обучаются» по множеству прототипов входных и выходных данных. Такие модели принято называть суррогатными моделями (1), (2).

Модели, основанные на данных (прототипах), по своей сути могут гарантированно использоваться только для входных данных, которые подобны входным данным прототипов, с помощью которых была построена модель. Тем самым, для нового множества прототипов модель должна быть либо перестроена (путем решения статистической задачи предсказания значений одной модели для некоторых входных данных по значениям другой модели для тех же входных данных), либо построена заново. Поэтому имеется настоятельная необходимость создания программных средств

автоматического (автоматизированного) создания моделей, основанных на данных.
В докладе описывается способ построения автоматических генераторов процедур, которые создают программные модули, реализующие базовые процедуры анализа и обработки данных, существенно зависят отданных и могут быть использованы как внешние библиотеки при разработке специализированных инженерных приложений.

2 Требования к генератору процедур

С основе каждой процедуры обработки данных лежит реализация одной или нескольких типовых математических задач. В настоящее время такие типовые математические задачи для заданного набора данных решаются «вручную» специалистом в области математических методов анализа данных. Для решения этих задач используются универсальные публично распространяемых математические программные пакеты, например, MATHLAB, MATHEMATICA, STATISTICA и др. При изменении исходного набора данных эти математические задачи должны решаться заново. Выбор метода решения задачи существенно зависит от структуры данных, для которой эта задача решается. Например, метод главных компонент является эффективным только в случае линейной структуры данных. Возможно также, что в данной предметной области существуют предметно-ориентированные процедуры снижения размерности, построенные с учетом «физики» данных (так называемые частные параметрические модели (5)).
Математик, знакомый с различными математическими методами решения задачи и понимающий, как влияет структура данных на качество той или иной процедуры, может «вручную» выбрать или построить достаточно эффективную процедуру размерности путем анализа структуры данных, комбинирования различных алгоритмов и проведения сравнительных вычислительных экспериментов.
Однако инженерные проектные команды не предполагают наличия в них математиков, а возможность оперативного привлечения «внешнего» математика затруднительна, а иногда и просто невозможна (например, в силу конфиденциального характера данных).
Подытоживая сказанное выше, можно сформулировать требования к генератору процедур.
Входными данными генератора является набор данных, поставляемый специалистом предметной области. Именно этот набор данных служит обучающим множеством при построении процедуры.
Для построения процедуры обработки данных **используются заранее известные классы функций и/или моделей**, которые могут включать в себя как универсальные модели, не зависящие от специфики данных, так и предметно-ориентированные модели, учитывающие «физику» данных.
Используемые классы моделей должны допускать комбинирование, то есть позволять последовательное применение одной модели обработки за другой. Это означает, что все модели имеют **единую спецификацию входов и выходов**.
Каждая модель может иметь **свой набор параметров**. Задавая значения этих параметров, пользователь может получить более точную настройку процесса генерации.
Выходом генератора процедур является программный модуль, который может быть использован как компонент при разработке специализированных инженерных приложений. Этот модуль реализуется как динамически подключаемая библиотека и может подключаться к приложению и активироваться без предварительной сборки или компиляции приложения.

3 Архитектура автоматического генератора процедур

В общем случае генератор процедур состоит из четырех компонентов.
Модуль импорта реализует функции загрузки входных данных.
Модуль генерации реализует следующие функции:

- выбор из существующего класса моделей обработки данных те, которые будут участвовать в генерации;
- задание параметров каждой отдельной модели;

- задание параметров процесса генерации;
- запуск генератора процедур.

Модуль экспорта формирует набор выходных файлов, который состоит из

- библиотеки, реализующей сгенерированную процедуру (например, библиотека может быть реализована в виде Win32 DLL;
- программный интерфейс приложений (API), предназначенный для вызова функций этой библиотеки из внешних систем;
- дополнительно может быть сгенерирован текстовый файл, содержащий формализованное описание параметров сгенерированной библиотеки.

Программный интерфейс сгенерированной процедуры должен содержать:

- описание процедуры, например, ее название, краткое описание алгоритма и др. Эта информация может быть включена в интерактивную справочную систему при подключении программного модуля;
- интерфейс по доступу к настраиваемым параметрам с наличием справочной информации о каждом параметре.

Ядро генератора содержит набор базовых программных модулей, которые согласованы по входным и выходными данным, то есть допускают последовательное применение одного модуля за другим к обучающему набору данных. Кроме того, каждый модуль обладает следующими свойствами:

- реализует частную процедуру обработки данных;
- имеет набор параметров и механизмы их чтения и модификации;
- может выдавать справочную информацию о себе, например, свой идентификатор или краткое описание реализованной процедуры;
- может выдавать справочную информацию о каждом из своих параметров, например, идентификатор параметра и его описание (тип, диапазон значений, значение по умолчанию и т.д.).

4 Пример реализации автоматического генератора процедур

Автоматический генератор процедур был разработан и успешно применяется для решения задачи сокращения размерности цифрового описания объекта.
Генератор реализует решение следующей математической задачи: по заданному набору данных, состоящего из N n-мерных векторов $\{X_1, X_2, \ldots, X_N\}$, необходимо выбрать размерность $m \leq n$ сжатого вектора и построить две процедуры:

- процедуру сжатия, преобразующую n-мерный вектор X в сжатый m-мерный вектор $\lambda = (X)$;
- процедуру восстановления, преобразующую m-мерный сжатый вектор λ в восстановленный n-мерный вектор $X^* = R(\lambda)$.

Качество этой пары процедур, примененных к исходному вектору X, определяется ошибкой восстановления - расстоянием $d(X, X^*) = |X - X^*|$ между исходным вектором X и восстановленным вектором $X^* = R(C(X))$, являющимся результатом последовательного применения процедур сжатия и восстановления. Для заданного набора данных, качество процедуры снижения размерности R определяется среднеквадратической ошибкой восстановления. В разработанном генераторе реализовано два независимых решения сформулированной задачи:

- при заданной точности (среднеквадратической ошибке восстановления) ϵ минимизируется размерность m сжатого вектора;
- при заданной размерности m сжатого вектора минимизируется среднеквадратическая ошибка восстановления.

Ядро генератора составили следующие классы алгоритмов:

- линейные алгоритмы снижения размерности, построенные на базе метода анализа главных компонентов;
- нелинейные алгоритмы снижения размерности, построенные на базе технологии искусственных нейронных сетей;
- класс частных параметрических моделей, разработанных для аналитического описания структур профилей.

Сравнительный Анализ автоматически сгенерированных процедур и стандартных процедур снижения размерности, основанных на методах Анализа Главных Компонент и технологии Искусственных нейронных сетей. Результаты сравнения приведены в таблице ниже.

Размерность сжатого вектора	ANN	PCA	Сгенерированная процедура
$m = 2$	4,11E-03	4,32E-03	3,06E-03
$m = 4$	1,96E-03	2,09E-03	1,89E-03
$m = 6$	1,36E-03	1,19E-03	1,10E-03
$m = 8$	9,65E-04	7,74E-04	6,63E-04
$m = 10$	7,15E-04	4,70E-04	4,64E-04

Вывод: Средние ошибки восстановления автоматически сгенерированных процедур равномерно меньше средних ошибок восстановления стандартных процедур снижения размерности, основанных на методах Анализа Главных Компонент и технологии Искусственных нейронных сетей.

Список литературы

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

[2] Бернштейн А.В., Кулешов А.П. Математические методы в когнитивных инженерных технологиях/ Обозрение прикладной и промышленной математики, сер. *Вероятность и статистика*. 2008, т. 15, № 3, с. 451 – 452

[3] Бернштейн А.В., Бурнаев Е.В., Дорофеев Е.А., Свириденко Ю.Н., Чернова С.С. О решении некоторых задач анализа данных, возникающих при построении адаптивных суррогатных моделей сложных объектов Научно-теоретический международный журнал *Искусственный интеллект*, Донецк, 2008, вып. 4, с. 40-48.

[4] Бернштейн А.В., Кулешов А.П. Когнитивные технологии в проблеме снижения размерности описания геометрических объектов. *Информационные технологии и вычислительные системы*. 2008, №2, с. 6 – 19.

[5] Иванова Е.П., Чернова С.С. Снижение размерности сложных геометрических объектов при наличии частных параметрических моделей. *Информационные технологии и вычислительные системы*, 2008, №4.

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