ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ
(Вычислительные системы)
2002 год Выпуск 171

УДК 519.31

СРАВНИТЕЛЬНЫЙ АНАЛИЗ РАСЧЕТНЫХ И ЭКСПЕРИМЕНТАЛЬНЫХ ЗАКОНОМЕРНОСТЕЙ[1]

А.Ю.Анохин, Н.Г.Загоруйко, А.Г.Пичуева

Введение

В работе [1] описаны экспериментальные данные, характеризующие изотопный состав контейнеров с отработанным ядерным топливом. Каждый контейнер описан пятимерным вектором процентного содержания пяти основных изотопов плутония. Количество контейнеров, хранящихся сейчас на складах, достигает несколько тысяч.

Данные по изотопному составу отработанного ядерного топлива можно получить и расчетным путем. Для известного типа реактора, его конструкционных особенностей и т.д. рассчитывается динамика образования элементов в реакторе и оценивается ожидаемый состав отработанного топлива. В настоящее время существует несколько расчетных моделей (см., например, [2,3]), каждая из которых дает сравнимые между собой результаты, но не имеющие достаточного экспериментального обоснования. В связи с этим возникает проблема верификации расчета нейтронно–физических характеристик реакторов.

Самым естественным путем решения этой проблемы является сравнение расчетных результатов с накопленными экспериментальными данными. Затруднение вызывает тот факт, что ошибки экспериментальных измерений (порядка ±1%) могут

[1]Работа выполнена при финансовой поддержке РФФИ, проект 02–01–00082.

быть сравнимыми с величиной различия результатов, получаемых разными методами. По этой причине простое сравнение расчетных и экспериментальных характеристик отдельных контейнеров не позволит сделать статистически достоверных выводов о преимуществах или недостатках различных методов расчета или оценить неопределенность получаемого результата.

Необходимо переходить к использованию некоторых интегральных характеристик, на значении которых несмещенные погрешности отдельных измерений сказываются в меньшей степени. В качестве таковых были выбраны характеристики регрессионных уравнений, описывающих зависимости между значениями пяти измеряемых изотопов на экспериментальных данных большого объема.

1. Предварительная обработка

В работе [1] отмечалось, что некоторые из строк таблицы экспериментальных данных содержали ошибки. В случае, когда ошибками поражены несколько элементов строки, для их обнаружения и исправления применялась программа ZET в режиме редактирования [4]. Этот алгоритм предназначен для прогнозирования значений пропущенных элементов в таблице (заполнение пробелов) и для редактирования (проверки) все таблицы или ее части.

В основе алгоритма ZET лежат три предположения.

Первое (гипотеза избыточности) состоит в том, что реальные таблицы имеют избыточность, проявляющуюся в наличии похожих между собой объектов (строк) и зависящих друг от друга свойств (столбцов). Если же избыточность отсутствует (как, например, в таблице случайных чисел), то предпочесть один прогноз другому невозможно.

Второе предположение (гипотеза локальной компактности) состоит в утверждении, что для предсказания пропущенного элемента a_{ij} нужно использовать не всю таблицу, а лишь ее "компетентную" часть, состоящую из элементов строк, похожих на строку i, и элементов столбцов, похожих на столбец j. Остальные строки и столбцы будут для данного элемента не информативными. Их использование лишь разрушало бы локальную

компактность подмножества "компетентных" элементов и ухудшало бы точность предсказания.

Третье предположение (гипотеза линейных зависимостей) заключается в том, что из всех возможных видов зависимостей между столбцами (строками) в алгоритме ZET используются только линейные зависимости. Если зависимости носят более сложный характер, то для их надежного обнаружения требуется такой большой объем данных, который в реальных задачах встречается не часто.

В работе алгоритма ZET можно выделить три этапа.

1. На первом этапе для данного пробела из исходной матрицы "объект-свойство", столбцы которой нормированы по дисперсии, выбирается подматрица "компетентных" строк и затем для этих строк — подматрица "компетентных" столбцов.

2. На втором этапе автоматически подбираются параметры в формуле, используемой для предсказания пропущенного элемента, при которых ожидаемая ошибка предсказания достигает минимума.

3. На третьем этапе выполняется непосредственное прогнозирование элемента по этой формуле.

Под "компетентностью" l-й строки по отношению к i-й строке понимается величина $L_{il} = r_{il} \cdot t_{il}$. Здесь $r_{il} = 1 - \rho_{il}$, ρ_{il} — Евклидово расстояние между i-й и l-й строками, а t_{il} — коэффициент комплектности, равный числу свойств, значения которых известны как для i-й, так и для l-й строки. Компетентная строка не должна иметь пробела в j-м столбце.

По указанию пользователя программа выбирает компетентную подматрицу любого размера в пределах от $2 * 2$ до $n * m$. Обычно используется подматрица, содержащая от трех до семи строк и столбцов.

В процессе предсказания значения пробела с использованием зависимостей между j-м и всеми остальными (k-ми) столбцами вырабатываются "подсказки" b_k. Для их получения используется уравнение линейной регрессии между j-м и k-м столбцами.

Если в подматрице было $(q + 1)$ столбцов, то затем q подсказок усредняются с весом, пропорциональным компетентности соответствующего столбца. В итоге получается прогнозная вели-

чина b_q, порожденная избыточностью, содержащейся в столбцах:

$$b_q = \frac{\sum_{k=1}^{q} b_k \cdot L_{jk}^{\alpha}}{\sum_{k=1}^{q} L_{jk}^{\alpha}}. \qquad (1)$$

Здесь α — коэффициент, регулирующий влияние компетентности на результат предсказания. При малых значениях α разница в компетентности сказывается мало, при больших α более компетентные столбцы влияют гораздо больше других. Выбор α и составляет суть этапа подбора формулы для прогнозирования: все известные элементы j–го столбца предсказываются при разных значениях α и затем выбирается такое значение α, при котором ошибка прогноза δ_j оказалась минимальной.

По формуле (1) с выбранным значением α делается прогноз b_q величины пропущенного элемента, а полученная при выборе α минимальная величина ошибки δ_j в дальнейшем принимается в качестве оценки ожидаемой ошибки заполнения пробела по столбцам.

Процедура заполнения пробела с использованием зависимости между i–й строкой и всеми s другими (l–ми) строками $(1, 2, \ldots, l, \ldots, s)$ аналогична вышеописанной и делается по формуле:

$$b_S = \frac{\sum_{l=1}^{S} b_l \cdot L_{il}^{\alpha}}{\sum_{l=1}^{S} L_{il}^{\alpha}}. \qquad (2)$$

Для выбора α здесь используются все известные элементы i–й строки и выбор делается при минимальном значении ошибки δ_i их прогнозирования.

Общий прогноз b'_{ij} значения пропущенного элемента b_{ij} — получается их усреднением с весом, обратно пропорциональным величине ожидаемой ошибки:

$$b'_{ij} = \left\{ \frac{b_q}{\varepsilon + \delta_j} + \frac{b_e}{\varepsilon + \delta_i} \right\} \frac{[\varepsilon + \delta_j][\varepsilon + \delta_i]}{2\varepsilon + \delta_j + \delta_i}. \qquad (3)$$

Здесь ε — константа, например, равная 0,01, введенная для предотвращения деления на 0.

Как отмечалось выше, оценка ожидаемой ошибки заполнения пробела (отклонения предсказанного значения от истинного) может быть получена в процессе подбора коэффициента α. О величине ожидаемой ошибки d_{ij} можно судить по ошибкам δ_i и δ_j предсказания известных элементов i-й строки и j-го столбца. Эксперименты показывают, что корреляция между средним значением этих ошибок $\delta^* = \frac{[\delta_j + \delta_i]}{2}$ и ошибкой d_{ij} всегда положительна.

Второй способ определения ожидаемой ошибки основан на оценке дисперсии "подсказок". Вычисляется дисперсия (dis) величин подсказок b_k и b_l, получаемых от всех k столбцов и l строк компетентной подматрицы. Большая дисперсия указывает на отсутствие устойчивой закономерной связи между элементом b_{ij} и другими элементами подматрицы, т.е. на отсутствие их компактности. Ясно, что в этих условиях рассчитывать на высокую точность предсказания величины b_{ij} не приходится. Эксперименты показали, что коэффициент корреляции между дисперсией dis и ошибкой предсказания d_{ij} достигает величины +0,7. Прогнозы ожидаемой ошибки заполнения по дисперсионному критерию оказались более надежными, чем по критерию, основанному на оценках ошибок δ.

Для различных прикладных задач были сделаны многочисленные модификации описанного выше базового алгоритма ZET, отличающиеся своим назначением и наборами разных режимов работы. Алгоритм ZET-R используется для обнаружения грубых ошибок в исходной таблице данных (так называемый режим "редактирования" таблиц). Для этого программа по очереди предсказывает все элементы таблицы и сравнивает результаты предсказания с фактически имеющимися данными. Если предсказанное значение совпадает с исходным, или мало отличается от него, то это означает, что этот элемент хорошо согласуется с закономерностями данной части таблицы данных. Если же обнаруживается большое расхождение, то выдается сигнал о необходимости проверки данного элемента. Если он отражает уникальный факт, выпадающий из общей закономерности, то его

истинность нужно подтвердить. Если же он отражает ошибку, то ее надо устранить. Таким путем удается обнаружить грубые ошибки или умышленные искажения отдельных элементов таблицы данных.

Именно этот режим редактирования таблиц был использован для обнаружения и исправления грубых ошибок в таблицах экспериментальных данных. После этого строились регрессионные уравнения, характеризующие связи между всеми пятью характеристиками материалов, содержащихся в контейнерах.

2. Регрессионный анализ

Для регрессионного анализа экспериментальных данных были использованы стандартные статистические пакеты программ. Были получены регрессионные уравнения для всех объектов таблицы данных и для подтаблиц, в которых отражены данные о контейнерах, заложенных в один и тот же год. Для каждого варианта исследовалась регрессионная зависимость каждого из пяти параметров от всех комбинаций из остальных четырех параметров: от каждого из четырех по одному, от всех их парных сочетаний, от всех троек и, наконец, от всех четырех вместе. При этом исследовались регрессии первого и второго порядков.

Рассчитав статистику для выборки с применением метода наименьших квадратов, находим зависимость, которая наилучшим образом аппроксимирует имеющиеся данные [5,6]. При этом минимизируется функция потерь $\sum(y_i - y_i(x))^2$, где y_i — отклик, $y_i(x)$ — предсказанное значение отклика, $i = \overline{1,n}$. Рассматриваемая линейная модель имеет вид $y = b + \sum\limits_{j=1}^{m} a_l \cdot x_{(j)} + \varepsilon$, где $y = (y_1, y_2, \ldots, y_n)'$ — отклик; а $x = (x_{(1)}, x_{(2)}, \ldots, x_{(j)}, \ldots, x_{(m)})$, где $x_{(j)} = (x_1, x_2, \ldots, x_n)'_{(j)}$ — наблюдаемые значения из выборки объема n.

Вид результата, получаемого для каждого варианта, показан на примере линейной регрессионной зависимости третьего параметра от трех других параметров (1-го, 2-го и 5-го) для подтаблицы, содержащей n объектов:

y = столбец 3, x = 1, 2, 5
ВЫВОД ИТОГОВ

Регрессионная статистика

Множественный R	MR
R-квадрат	R_h
Нормированный R-квадрат	$H - R_h$
Стандартная ошибка	C_0
Наблюдения	n

Здесь

множественный R — это множественный коэффициент корреляции между истинным y и предсказанным значением отклика $y(x)$, мера их линейной зависимости;

R-квадрат $(R)^2$ — коэффициент детерминированности, который показывает, насколько хорошо линейное уравнение описывает реальные данные (насколько предсказание по регрессионной модели лучше, чем по среднему значению отклика $\overline{y}$);

нормированный R-***квадрат*** — также показатель линейной зависимости, он используется при сравнении моделей с разным числом степеней свободы;

стандартная ошибка — среднеквадратическая ошибка для оценки отклика;

наблюдения — объем выборки n.

Вариация	Степени свободы (df)	Сумма квадратов (SS)	Средний квадрат (VS)	F-отношение	Значимость F
Регрессия	m	SS_R	$MS_R(\sigma_{R^2})$	$\frac{MS_R}{MS_F}$	Significance F
Остаток	$n-m-1$	SS_R	$MS_R(\sigma_{R^2})$		
Итого	$n-1$	SS_R-SS_F			

Здесь

степени свободы df используются при вычислении параметров дисперсионного анализа и при нахождении обратного t-распределения и обратного F-распределения;

регрессионная сумма квадратов SS_R вычисляется как $\sum(y_i(x) - \overline{y})^2$, где $\overline{y} = \frac{1}{n}\sum y_i$;

остаточная сумма квадратов SS_E (сумма потерь) вычисляется как $\sum(y_i - y_i(x))^2$;

средний квадрат, обусловленный остаточной вариацией MS_E, представляет собой оценку дисперсии остатков σ_E^2;

средний квадрат, обусловленный регрессией MS_R, представляет собой оценку σ_R^2 вариации, обусловленной регрессией.

F–наблюдаемое значение распределения и значимость F. Оценить значимость уравнения регрессии — значит установить соответствует ли математическая модель, выражающая зависимость между y и x_j экспериментальным данным. Для оценки значимости вычисляют статистику F–наблюдаемое. Затем по таблицам находят значение F–критическое, которое соответствует статистике Фишера с уровнем значимости 0,05 и m, $(n - m - 1)$ степеням свободы. Если F–наблюдаемое больше F–критического, то регрессионное уравнение считают значимым. Кроме того, при таком результате сравнения значимым считается множественный коэффициент корреляции R. Значимость F показывает вероятность отвергнуть предложение об адекватности линейной модели, когда она верна.

Y–пересечение	Коэффициенты	Стандартная ошибка	*t*–статистика	*P*–значение	Нижние 95%	Верхние 95%
	a_0	s_0	t_0	P_0	L_0	H_0
1	a_1	s_1	t_1	P_1	L_1	H_1
2	a_2	s_2	t_2	P_2	L_2	H_2
3	a_3	s_3	t_3	P_3	L_3	H_3

Здесь

коэффициенты — это коэффициенты в исследуемом линейном уравнении $e(x) = b + \sum_{j=1}^{m} a_j \cdot x_{(j)}$;

стандартная ошибка — среднеквадратическая ошибка вычисления коэффициентов;

нижние 95% и верхние 95% — 95%–е доверительные интервалы для коэффициентов регрессионного уравнения.

t–статистика и *P–значение*. В рамках линейной теории множественной регрессии оценивается значимость каждой переменной при помощи t–критерия. Для оценки значимости вычисляют статистику t–наблюдаемое. Затем по таблицам находят значение t–критическое, которое соответствует статистике Стьюдента с у.з. 0,05 и $(n - m - 1)$ степенями свободы. Если абсолютная величина t–наблюдаемого больше t–критического, то соответствующая переменная является важной, т.е. статистически значимой, а предложение о равенстве этой переменной нулю отвергается.

Анализ полученных результатов показал, что регрессии второго порядка описывают эмпирические зависимости лучше, чем регрессии первого и третьего порядков. Вместе с тем, разница в точности между линейными и квадратичными регрессиями незначительна. По этой причине в дальнейшем анализировались линейные регрессии.

Как и ожидалось, зависимости между отдельными параметрами сильно отличаются друг от друга. Есть пары параметров, где каждый из них предсказывается по второму с высокой точностью. И есть пары связанные друг с другом слабой зависимостью.

Естественно, с ростом размерности регрессионного уравнения точность предсказания целевого параметра увеличивается.

Стандартные наибольшие ошибки получаются для тех параметров, абсолютная величина которых минимальна. Это подтверждает тот известный факт, что изотопы с малой процентной долей в массе контейнера измеряются с наибольшей относительной погрешностью. Отсюда следует, что при сравнении экспериментальных и расчетных данных следует обращать внимание в первую очередь на изотопы с большой процентной долей.

4. Сравнение экспериментальных и теоретических регрессий

Для тех же наборов объектов (контейнеров) были получены расчетные изотопные составы по специальным программам (например, TVS–M, SCALE, Apollo, MCU и др.). В данных

программах реализовано решение систем дифференциальных уравнений вида:

$$\left.\begin{aligned}
\frac{dN_1}{dt} &= -\lambda_1 N_1;\\
\frac{dN_2}{dt} &= -\lambda_1^c N_1 - \lambda_2 N_2;\\
\frac{dN_3}{dt} &= -\lambda_2^c N_2 - \lambda_3 N_3;\\
&\cdots\cdots\cdots\\
\frac{dN_n}{dt} &= -\lambda_{n-1}^c N_{n-1} - \lambda_n N_n;
\end{aligned}\right\}$$

где λ — константа, определяемая свойствами каждого изотопа; N_n — число ядер n–го изотопа. Величина $\lambda_{n-1}^c N_{n-1}$ — скорость захватов нейтронов ядрами типа $n-1$, равная скорости образования ядер типа n, а величина $\lambda_n N_n$ — скорость убыли ядер типа n за счет всех возможных процессов (например, распад, деление и т.д.).

По этим расчетным данным строились регрессионные уравнения и определялись коэффициенты регрессий (назовем их теоретическими). Сравнение теоретических и экспериментальных коэффициентов регрессий осуществлялось сравнением полученных коэффициентов. Оценочное значение неопределенности теоретических коэффициентов достигает нескольких процентов.

Результаты сравнений показали совпадение теоретических и экспериментальных результатов в пределах до 10%. Однако полученное значение несовпадения достаточно велико и может быть сокращено, если дополнить экспериментальные результаты условиями проведения экспериментов по каждому исследуемому вектору изотопов.

З а к л ю ч е н и е

Результаты анализа показали возможность использования результатов промышленного эксперимента для верификации нейтронно–физических расчетов. Однако получение более точных и достоверных результатов может быть обеспечено увеличением сведений по проведенным экспериментам (как, из чего и при

каких условиях получены эти данные для каждого контейнера). Получение такой информации является задачей дальнейших исследований.

Литература

1. АНОХИН А.Ю., БОРИСОВА И.А., ЗАГОРУЙКО Н.Г. Классификация образцов отработанного ядерного топлива// Настоящий сб. – С. 3–10.

2. ГЕРАСИМОВ А.С., ЗАРИЦКАЯ Т.С., РУДИК А.П. Справочник по образованию нуклидов в ядерных реакторах. – М.: Энергоатомиздат, 1989.

3. КРУГЛОВ А.К., РУДИК А.П. Искусственные изотопы и методика расчета их образования в ядерных реакторах.– М.: Атомиздат, 1977.

4. ЗАГОРУЙКО Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд.ИМ СО РАН, 1999. – 270 с.

5. ДРЕЙПЕР Н., СМИТ Г. Прикладной регрессионный анализ. – М.: Финансы и статистика, 1986.

6. ЛЕМАН Э. Проверка статистических гипотез. – М.: Наука,1979.– 408 с.

7. ХОЛЛЕНДЕР М., ВУЛФ Д. Непараметрические методы статистики. – М.: Финансы и статистика,1983. – 518 с.

Поступила в редакцию
29 октября 2002 года.