УДК 681.513

О.В. Бабак, А.Э. Татаринов

Об одном подходе к решению некорректных задач идентификации физических объектов при разведочном анализе данных

Рассмотрен новый эвристический подход к решению некорректно поставленных задач идентификации физических объектов при разведочном анализе данных. Показано, что для априори известного характера изменения функции отклика возможен объективный синтез мультипликативных обобщенных переменных, позволяющий свести некорректную задачу к корректной с сохранением ее физической сущности.

A new heuristic approach to solving ill-posed problems of the identification of physical objects in the exploratory data analysis is considered. It is shown that it is possible to reduce the ill-posed problem to the well-posed problem with preservation of the physical nature of the problem by means of synthesizing multiplicative generalized variables for the response function with a priori known character of its change.

Розглянуто новий евристичний підхід до розв'язання некоректно поставлених задач ідентифікації фізичних об'єктів при розвідувальному аналізі даних. Показано, що для апріорі відомого характеру зміни функції відгуку можливий об'єктивний синтез мультиплікативних узагальнених змінних, що дозволяє звести некоректну задачу до коректної з збереженням її фізичної сутності.

Введение. В различных предметных областях при разведочном анализе данных (РАД) [1] возникают задачи идентификации объектов в условиях неполных наблюдений. Дело в том, что получение выборок достаточно большого объема, особенно в производственных условиях, представляет почти всегда трудноразрешимую проблему. Объясняется это тем, что проведение необходимых экспериментов не только связано со значительными, а иногда невосполнимыми материальными затратами, но в отдельных случаях просто невозможно, например, из-за серьезного нарушения технологического режима. В этой связи сокращение числа экспериментов при РАД имеет важное практическое значение и требует поиска эффективных путей решения, возникающих при этом некорректных задач идентификации объектов.

Основная цель РАД заключается в создании моделей и методов, позволяющих анализировать многомерные данные с помощью их отображения в пространство низкой размерности. В общем случае решается задача перехода к компактному описанию исходных данных при возможно более полном сохранении существенных аспектов информации, содержащихся в них. При этом отображение многомерных данных в пространстве низкой размерности может быть как линейным, так и нелинейным.

В статье предлагается новый эвристический подход к решению некорректных задач идентификации объектов по данным физической природы. К таким данным в первую очередь можно отнести показания различного рода приборов, осуществляющих прямые измерения температуры, давления, силы тока, напряжения, веса, концентрации, расхода и других физических величин. Предлагаемый подход, по-видимому, в целом не может быть распространен на решение аналогичных задач при обработке данных медицинской, биологической, социальной, экономической, экологической и другой нефизической природы, получение которых имеет совершенно иную специфику.

В [2] приведено понятие физических и нефизических моделей. Физические модели содержат минимальное число постоянных (параметров) и соответствуют теории того или иного физического процесса. Нефизические модели, имея не минимизированное число параметров, не поддаются простой физической интерпретации и характерны тем, что получаются при помощи алгоритмов самоорганизации. Лишенные физического смысла такие модели, представляя собой просто интерполяционные формулы, часто обладают низкими экстраполяционными свойствами и малопригодны для управления и косвенных измерений.

Покажем, что в ряде случаев линейная по параметрам a и независимым переменным (факторам) x полиномиальная модель с зависимой переменной y

$$y = a_0 + \sum_{i=1}^{n} a_i x_i , \qquad (1)$$

где n – число факторов и a – оценки коэффициентов, может до известной степени отображать истинную физическую модель процесса.

Пусть известна истинная физическая мультипликативная модель процесса

$$y = \frac{x_1}{x_2} \text{ (например, закон Ома).} \qquad (2)$$

Устанавливая некоторые пределы изменения факторов x_1 и x_2, по ней можно определить ряд значений y_j, $j = \overline{1,l}$ и получить выборку $\{x_{ij}, y_j\}, i = \overline{1,2}, j = \overline{1,l}$. Нетрудно показать, что восстанавливая по ней зависимость (1), находим ее в виде

$$y = a_0 + a_1 x_1 - a_2 x_2, \qquad (3)$$

где знаки при оценках коэффициентов a_1 и a_2 – направления составляющих градиента функции отклика. Сравнивая истинную физическую модель с полученной, отметим, что в физическом отношении они подобны. Действительно, при увеличении (уменьшении) значения x_1 при заданном значении x_2 величина y увеличивается (уменьшается), а при увеличении (уменьшении) x_2 при заданном значении x_1 величина y уменьшается (увеличивается). Несмотря на то что модели (2) и (3) в определенных пределах изменения x_1 и x_2 обладают сравнимой точностью, модель (3) является не физической, а близкой ей по физической сущности.

Из рассмотренного простого примера следует важный вывод о том, что поскольку линейная модель (1) содержит информацию о направлении составляющих градиента функции отклика, то она несет информацию и о возможном характере нелинейности функции, отображающей истинную физическую модель. Тогда при условии аналитичности этой функции гарантировано существование подобласти пространства факторов, в которой линейная квазифизическая модель адекватна нелинейной истинной физической модели и наоборот. Отметим, что нелинейная истинная физическая модель необязательно может быть мультипликативной. Полученный вывод естественным образом распространяется и на случай, когда факторное пространство $n > 2$. Такие результаты, а также знание поведения функции отклика y от того или иного приращения значений x были положены в основу предлагаемого эвристического подхода. Рассмотрим только линейные функции, у которых емкость класса $h = n + 1$.

Постановка некорректно поставленной задачи

Пусть имеется конечное число данных физической природы x, отождествляемых с вектором наблюдений y, представленных в виде выборки

$$\{x_{ij}, y_j\}, i = \overline{1,n}, j = \overline{1,l}, \forall x_{ij} > 0, \qquad (4)$$

причем $l < (n+1)$. Все факторы x значимы и поэтому удаление отдельных из них недопустимо.

Априори известно влияние приращения значений факторов x на характер приращения функции отклика y. Требуется решить некорректно поставленную задачу идентификации объекта, имеющего физическую природу, путем построения модели, близкой ему по физической сущности.

Решение задачи

В [3] предложен алгоритм, позволяющий формализовать решение поставленной некорректной задачи, сущность которого состоит в следующем. Поскольку система нормальных уравнений $X_н \cdot a = y_н$, полученных в соответствии с выборкой (4), является вырожденной, то можно получить ее псевдорешение с помощью минимизации функционала

$$M_\alpha(a, y) = \| X_н \cdot a - y_н \|^2 + \alpha \| a \|^2, \alpha > 0, \qquad (5)$$

где α – неопределенный множитель Лагранжа, по невязке

$$\overline{\delta_н} = \frac{1}{l_н} \sum_{j=1}^{l_н} (X_н \cdot a_\alpha - y_{jн})^2, \; l_н = (n+1).$$

В результате минимизации (5) по вектору $a = (a_i)$, $i = \overline{0,n}$ каждый диагональный элемент C_{ij} главной диагонали квадратной матрицы $\det X_н$ вырожденной системы нормальных уравнений $X_н \cdot a = y_н$ представляет собой сумму $(C_{ij} + \alpha)$, $j = \overline{1,l_н}$, и множитель α однозначно находится путем подбора. Однако результатом реализации

алгоритма является модель, которая не отвечает физической сущности исследуемого объекта и может быть интерпретирована только как интерполяционная формула (см. Приложение). Более того, за ее точность и пригодность для практических целей трудно ручаться в связи с неопределенностью выбора значения $\overline{\delta_и}$.

Устранить эти недостатки можно с помощью предлагаемого эвристического подхода к построению алгоритма решения поставленной задачи, основанного на методе главных (генеральных) обобщенных переменных (МГОП) [4–6], в основе которого лежат три положения.

• Если функция отклика физического объекта $y = f(x)$ в заданной окрестности некоторой точки аналитична, то восстановленная линейная по параметрам a и переменным x зависимость $y = f(x)$ является моделью объекта, близкой по физической сущности, поскольку содержит информацию о направлении составляющих градиента $y = f(x)$.

• Если $y = f(x)$ в заданной окрестности некоторой точки аналитична, то всегда существуют пределы изменения x, в которых восстановленная $y = f(x)$ близка по оценке остаточной суммы квадратов любой восстановленной линейной по параметрам и нелинейной по переменным функции.

• Если $y = f(x)$ в заданной окрестности некоторой точки аналитична, то любая восстановленная линейная по параметрам a и нелинейная по обобщенным переменным v, синтезированным с учетом направления составляющих градиента $y = f(x)$, также является моделью объекта, близкой ему по физической сущности.

С учетом указанного в основе МГОП лежит идея о том, что необходимую информацию о влиянии факторов на функцию отклика, а также о характере нелинейной зависимости, скрытой в данных, можно получить из линейного по параметрам и независимым переменным полинома (1). Важной особенностью (1) является то, что он содержит оценки о направлении составляющих градиента функции отклика ∇y. Напомним, что его составляющими служат частные производные функции отклика, оценками которого являются коэффициенты a_i, $i = \overline{1,n}$

$$\nabla y = \sum_{i=1}^{n} \frac{\partial y}{\partial x_i} z_i \ ,$$

где z_i – единичный вектор, параллельный соответствующей координатной оси.

Частные производные характеризуют изменение функции y по каждой независимой переменной x в отдельности. Образованный с их помощью вектор ∇y дает общее представление о поведении функции в окрестности выбранной точки как в случае линейности, так и ее нелинейности. Практическое выражение данной идеи состоит в синтезе генеральной обобщенной переменной (ГОП), которая может быть представлена мультипликативной функцией v

$$v = \prod_{i=1}^{n} x_i^{p_i} \ , \qquad (6)$$

где p_i – величина, принимающая значение ±1 в зависимости от знака при соответствующей оценке градиента a_i функции отклика (1), восстановленной по эмпирическим данным. При этом замечательным свойством функций (1) и (6), как было показано во введении на очень простом примере, представляется то, что они в определенном смысле оказываются подобными. Это свойство, гарантирующее непротиворечивость приращений зависимых величин y и v от приращения любой независимой величины x, дает все основания утверждать, что синтез v в определенном смысле оптимален. Если же на эвристическом уровне известны направления составляющих градиента линейной функции, описывающей объект физической природы, то синтез v может быть осуществлен без предварительного восстановления линейной зависимости (1). Справедливо следующее утверждение.

Утверждение. Пусть задана линейная функция (1) (или на эвристическом уровне направления составляющих градиента такой функции), тогда результатом ее нелинейной аппроксимации (о задаче нелинейной аппроксимации см. [7]) будет зависимость

$$y = a_0' + \sum_{i=1}^{n'} a_i' v_i,\ n' < n, \qquad (7)$$

линейная по параметрам a_0', a_i' и по переменным v_i, представляющим собой мультипликативные функции, синтезированные по аналогии с функцией (6). При этом оценки коэффициентов при v_i принимают знак «+».

Справедливость утверждения определяется тем, что вклад в значение функции отклика y любой из частных функций v_i, $i = \overline{1, n'}$ не может быть отрицательным в силу свойства, гарантирующего непротиворечивость приращения зависимых величин y и v от приращения любой независимой величины x.

Отметим, что по существу МГОП является процедурой проецирования исходных многомерных данных в спрямляющее пространство гораздо меньшей размерности, в предельном случае одномерной. В нашем случае неопределенная система условных уравнений в исходной задаче становится определенной или переопределенной, т.е. исходная выборка (4) преобразуется к виду

$$\{v_{ij}, y_j\}, i = \overline{1, n'}, l \geq (n'+1), \qquad (8)$$

и известными методами восстанавливается зависимость (7).

Таким образом, алгоритм решения поставленной задачи включает в себя три этапа.

Этап 1. На основе априорной информации о физической сущности объекта синтез различных вариантов набора функций v_i, $i = \overline{1, n'}$.

Этап 2. Преобразование исходной выборки (4) в выборку (8).

Этап 3. Восстановление зависимости (7).

Полученная модель (7) в дальнейшем может быть использована для различных целей РАД, в частности, для выяснения степени значимости той или иной переменной x или их комплекса при управлении объектом и т.д.

Оценка качества решения задачи

В работе [8] на основе неравенства П.Л. Чебышева, справедливого для всех законов распределения случайных величин, доказана теорема, смысл которой состоит в следующем. С вероятностью (1 – η) (η – риск) можно утверждать, что функционал среднего риска $I(a)$ находится в пределах

$$I_э(a) - \chi \leq I(a) \leq I_э(a) + \chi, \qquad (9)$$

где $I_э(a) = \frac{1}{l}\sum_{j=1}^{l}(y_j - F(x_j, a))^2$ – функционал эмпирического риска, χ – некоторое действительное число. При этом показано, что величина $\chi = \tau_a \sqrt{\frac{\ln N - \ln \frac{\eta}{2}}{2l}}$, $\chi = \tau_a \cdot \chi_0$, где оценка $N = 2^h$, $h = n' + 1$, τ_a – оценка возможного «выброса» ($\sup (y - F(x, a))^2 \leq \tau_a$). Полагая, что $\tau_a \approx \approx I_э(a)$ из правой части неравенства (9), получим оценку $I(a) = I_э(a)(1 + \chi_0)$. Используя ее, можно найти процентное отношение величины эмпирического риска к величине среднего риска

$$I_э(a)\% = \frac{100}{(1+\chi_0)}. \qquad (10)$$

Соотношение (10) – оценка качества решения некорректной задачи идентификации объекта физической природы, из чего следует, что чем больше значение $I_э(a)$%, тем ближе оно находится к величине $I(a)$ и, соответственно, выше качество решения.

Приложение

Пример. Задана некоторая выборка $\{x_{ij}, y_j\}$, $i = \overline{1, 8}$, $j = \overline{1, 30}$, $h = 9$, $l = 30 \leq 10 \cdot h$ в виде табл. 1П.

Таблица 1П. Исходная выборка данных

№	x_1	X_2	x_3	x_4	x_5	x_6	x_7	x_8	y
1	**3,96**	**4,00**	**5,20**	**2,34**	**3,63**	**5,08**	**1,55**	**4,23**	**347,42**
2	**3,99**	**4,10**	**5,73**	**2,77**	**3,40**	**4,08**	**1,75**	**3,91**	**348,53**
3	**4,68**	**3,72**	**5,05**	**2,72**	**3,38**	**4,38**	**1,90**	**3,85**	**354,55**
4	**4,31**	**4,11**	**5,09**	**2,35**	**4,67**	**4,87**	**1,50**	**3,98**	**355,31**
5	**5,74**	**3,76**	**5,71**	**2,73**	**3,81**	**4,47**	**1,29**	**4,97**	**367,20**
…	…	…	…	…	…	…	…	…	…	…
28	**5,79**	**3,76**	**4,82**	**2,38**	**4,71**	**4,19**	**2,73**	**4,76**	**390,14**
29	**4,36**	**4,01**	**5,00**	**2,45**	**4,62**	**4,59**	**2,85**	**4,17**	**384,43**
30	**6,05**	**3,63**	**5,35**	**3,00**	**3,67**	**4,56**	**2,00**	**4,23**	**387,07**

Нормируя x_i $\left(x_{iн} = \frac{x_i}{x_{i\max}}\right)$, восстанавливаем методом наименьших квадратов (МНК) линейную зависимость (1) для того, чтобы «под-

смотреть» направление составляющих ее градиента (т.е. знаки при оценках коэффициентов регрессии). Получаем

$$y = 53,08 + 57,10 \cdot x_1 - 33,34 \cdot x_2 + \\ + 102,79 \cdot x_3 + 54,77 \cdot x_4 + 58,87 \cdot x_5 + \\ + 95,87 \cdot x_6 + 87,06 \cdot x_7 - 8,52 \cdot x_8 . \quad (1П)$$

Соотношение (1П) отвечает физической сущности объекта. Выберем в выборке произвольно пять точек: точки 6–10. При этом точки 1–5 и 11–30 будут служить в качестве проверочной выборки. Для анализа полученных результатов используем величину $\gamma = \sqrt{\bar{\delta}}$, где $\bar{\delta} = \frac{1}{l}\sum_{j=1}^{l}(y_j - y_{wj})^2$ – среднее арифметическое остаточной суммы квадратов. Восстановление зависимости по выборке $\{x_{ij}, y_j\}$, $i = \overline{1,8}$, $j = \overline{1,5}$ является некорректно поставленной задачей. Рассмотрим три совершенно произвольно взятых варианта ее решения на основе МГОП. В каждом из трех вариантов приводим исходную выборку к ситуации, когда $i = \overline{1,3}$, $j = \overline{1,5}$, т.е. когда задача восстановления зависимости становится корректно поставленной.

Поскольку известно направление составляющих градиента функции отклика (1П), осуществляем синтез ГОП:

Вариант 1

$$v_{1j} = x_{1j} \cdot x_{5j}, v_2 = \frac{x_{3j} \cdot x_{7j}}{x_{2j}}, v_3 = \frac{x_{4j} \cdot x_{6j}}{x_{8j}}.$$

Вариант 2

$$v_{1j} = x_{3j} \cdot x_{7j}, v_2 = \frac{x_{4j} \cdot x_{6j}}{x_{2j}}, v_3 = \frac{x_{1j} \cdot x_{5j}}{x_{8j}}.$$

Вариант 3

$$v_{1j} = x_{4j} \cdot x_{6j}, v_2 = \frac{x_{1j} \cdot x_{5j}}{x_{2j}}, v_3 = \frac{x_{3j} \cdot x_{7j}}{x_{8j}}.$$

Соответственно, восстановленные с помощью МНК зависимости имеют вид:

Вариант 1

$$y = 267,47 + 49,01 \cdot \nu_1 + 43,70 \cdot \nu_2 + 60,41 \cdot \nu_3 ;$$

Вариант 2

$$y = 239,01 + 82,03 \cdot \nu_1 + 71,88 \cdot \nu_2 + 53,12 \cdot \nu_3 ; \quad (2П)$$

Вариант 3

$$y = 300,55 + 39,37 \cdot \nu_1 + 32,21 \cdot \nu_2 + 37,44 \cdot \nu_3 .$$

Все они отвечают физической сущности объекта, поскольку величины v синтезированы на основе (1П). Отметим, что оценки коэффициентов при v у полученных зависимостей принимают знак «+».

Алгоритм А.Н. Тихонова реализуем, подбирая неопределенные множители Лагранжа α. При этом в качестве невязки $\bar{\delta}_1$ используем значения $\bar{\delta}$, полученные на пяти точках (6–10) выборки для трех вариантов решения некорректной задачи с помощью алгоритма, построенного на основе МГОП. В данном случае при $\alpha_1 = 0,0016$, $\alpha_2 = 0,0006$ и $\alpha_3 = 0,0034$, соответственно, имеем псевдорешения:

Вариант 1

$$y = 80,61 + 26,26 \cdot x_1 + 46,62 \cdot x_2 + 101,57 \cdot x_3 + \\ + 21,29 \cdot x_4 + 38,18 \cdot x_5 + 60,45 \cdot x_6 + 71,24 \cdot x_7 - 0,04 \cdot x_8 .$$

Вариант 2

$$y = 81,08 + 26,36 \cdot x_1 + 46,48 \cdot x_2 + 102,14 \cdot x_3 + \\ + 19,96 \cdot x_4 + 38,00 \cdot x_5 + 62,01 \cdot x_6 + 71,68 \cdot x_7 - 1,44 \cdot x_8 . \quad (3П)$$

Вариант 3

$$y = 79,92 + 26,18 \cdot x_1 + 46,82 \cdot x_2 + 100,67 \cdot x_3 + \\ + 23,10 \cdot x_4 + 38,51 \cdot x_5 + 58,35 \cdot x_6 + 70,52 \cdot x_7 + 1,94 \cdot x_8 .$$

Сравнивая их с моделью (1П), близкой по физической сущности объекту, отметим, что они отличаются от нее по знакам оценок коэффициентов при x_2 (1, 2 и 3 псевдорешение) и при x_8 (3 псевдорешение). Поэтому соотношения (3П), являясь интерполяционными формулами, не отвечают физической сущности объекта.

Полученные результаты для сравнения отображены в таблице 2П, где v – алгоритм МГОП, а T – алгоритм А.Н. Тихонова.

Таблица 2П. Сводная таблица результатов

Алгоритм	Кол-во точек	Оценки	Вариант 1	Вариант 2	Вариант 3
v	5	$\bar{\delta}_1$	0,09	0,01	0,35
		γ_1	0,31	0,11	0,59
	25	$\bar{\delta}_2$	88,32	37,97	62,30
		γ_2	9,40	6,16	7,89
T	5	$\bar{\delta}_1$	0,09	0,01	0,35
		γ_1	0,31	0,11	0,59
	25	$\bar{\delta}_2$	58,86	59,00	58,24
		γ_2	7,67	7,70	7,63

Анализ результатов показывает, что для всех трех вариантов они сопоставимы. Результаты оценки качества исходной выборки, а также для трех вариантов решения некорректно поставленной задачи представлены в таблице 3П.

Таблица 3П. Результаты оценки качества вариантов решений

	Исходная выборка	Выборка ГОП, варианты		
		1	2	3
η	0,05	0,05	0,05	0,05
l	30	5	5	5
h	9	4	4	4
N	512	16	16	16
τ_a	14,04	0,25	0,02	1,14
χ	5,71	0,20	0,02	0,92
$I_э(a)$	4,21	0,09	0,01	0,35
$I(a)$	9,92	0,30	0,03	1,27
$I_э(a)\cdot 100 / I(a)$	42,42%	31,71%	40,72%	27,74%

Анализ результатов, приведенных в таблицах 2П и 3П, показывает, что наиболее предпочтительным является второй вариант решения задачи.

Заключение. При автоматизации технологических процессов нередко в силу тех или иных причин возникает проблема РАД, связанная с сокращением экспериментов и идентификацией физических объектов по недостаточному числу данных. В свою очередь это приводит к необходимости решения некорректно поставленных задач идентификации физических объектов. Известный метод решения таких задач путем подбора псевдорешения [3] приводит к построению модели в виде интерполяционной формулы, не отвечающей физической сущности исследуемого объекта. Показано, что если априори известен характер изменения функции отклика от приращения той или иной независимой переменной, то возможен эвристический подход к решению таких задач. При этом важной особенностью полученной модели является то, что она оказывается близкой по физической сущности исследуемому объекту. Таким образом, достигается главная цель РАД – построить некоторую статистическую модель отбора нужных данных, которую, естественно, дальше нужно верифицировать.

1. *Прикладная* статистика: Классификация и снижение размерности: Справочное изд. / Под ред. С.А. Айвазяна. – М.: Финансы и статистика, 1989. – 426 с.
2. *Ивахненко А.Г.* Индуктивный метод самоорганизации моделей сложных систем. – Киев: Наук. думка, 1981. – 286 с.
3. *Тихонов А.Н.*, *Арсенин В.Я.* Методы решения некорректных задач. – М.: Наука, 1986. – 283 с.
4. *Бабак О.В.* Об одном принципе самоорганизации математических моделей // Проблемы управления и информатики. – 2001. – № 2. – С. 98–107.
5. *Бабак О.В.* Новый подход к решению задач идентификации объектов управления по эмпирическим данным // УСиМ. – 2001. – № 2. – С. 25–31.
6. *Бабак О.В.* Решение некоторых задач обработки данных на основе метода генеральной обобщенной переменной // Проблемы управления и информатики. – 2002. – № 6. – С. 78–91.
7. *Коллатц Л.* Функциональный анализ и вычислительная математика. – М.: Изд. Мир, 1969. – 447 с.
8. *Алгоритмы* и программы восстановления зависимостей / Под ред. В.Н. Вапника. – М.: Наука, 1984. – 816 с.

Поступила .30.03.2010

Тел. для справок: (044) 526-4187, 502-6337 (Киев)
E-mail: *dep*175@*irtc.org.ua*