Математика

УДК 519.6
DOI: 10.22363/2312-9735-2017-25-4-323-330

Индуктивное моделирование объектов и явлений методом группового учёта аргументов: недостатки и способы их устранения

М. Ю. Дьячков

Кафедра нелинейного анализа и оптимизации
Российский университет дружбы народов
ул. Миклухо-Маклая, д. 6, Москва, Россия, 117198

Представлены оригинальные результаты исследования эффективного вычислительного метода — метода группового учёта аргументов. Выявлены и систематизированы ключевые недостатки на каждой значимой процедуре классического алгоритма, а также представлены способы их устранения, в том числе авторские модификации. В частности, предложено использование дисперсии и её оценки (критерий Фишера) в качестве оценки точности полученного результата, дополнительного «внутреннего» критерия оценки адекватности модели в различных тестах при фиксации исходных данных и изменении характеристик алгоритма, а также определения оптимальной сложности модели. Для решения проблемы сходимости классического алгоритма было предложено использование методов дисперсионного, факторного и корреляционного анализов для исключения неинформативных признаков, модификации критерия остановки алгоритма. Предложено использование регуляризирующих функционалов для разрешения проблемы мультиколлинеарности входных признаков и повышения устойчивости полученной модели и др. Разработан комплекс программ компьютерного моделирования, реализующий модифицированный эффективный алгоритм метода группового учёта аргументов с рассмотренными авторскими модификациями, а также методами дисперсионного анализа, корреляционного анализа, факторного анализа, элементы регрессионного анализа и др. Проведённые исследования и полученные практические результаты могут стать основой для разработки с применением современных технологий Machine Learning и Data Science автоматизированной системы компьютерного моделирования, интеллектуального анализа и обработки данных.

Ключевые слова: математическое моделирование, индуктивное моделирование, метод группового учёта аргументов, эффективный алгоритм МГУА, адекватная модель, комплекс программ

1. Введение

Математическое моделирование стало неотъемлемой частью построения и исследования функционирования сложных систем. В условиях неполноты информации активно используются методы индуктивного моделирования, позволяющие последовательно строить модели возрастающей сложности непосредственно по выборке данных без привлечения дополнительной априорной информации в фиксированном классе функций, предназначенные для функционального описания входо-выходных характеристик систем.

Среди индуктивных методов математического моделирования выделяется эффективный вычислительный метод — метод группового учёта аргументов (МГУА) [1–3]. Его основой являются алгоритм массовой селекции (принцип самоорганизации моделей), теорема Геделя о неполноте, принцип сохранения свободы выбора Габора [2, 3]. МГУА относится к методам анализа данных (Data Science), успешно применяется для решения задач моделирования, прогнозирования, распознавания образов и др. Его алгоритмы позволяют находить взаимозависимости и закономерности, обнаруживать новые знания, неявно отражённые в данных, и представлять их

Статья поступила в редакцию 31 марта 2017 г.

в виде адекватной математической модели оптимальной сложности, представленной в явном функциональном виде [1–3]. Проверка соответствия математической модели исходным данным, то есть её адекватность, оценивается точностью прогноза и сложностью структуры, а также согласованностью обнаруженных знаний результатам исследования.

Однако классический многорядный алгоритм МГУА не лишён недостатков, которые могут не позволить корректно решить поставленную задачу. Некоторые из них могут быть устранены за счёт приведённых в статье ранее предложенных исследователями научно обоснованных решений, а также авторских модификаций, использование которых может варьироваться при решении различных классов задач [1–3]. На каждой значимой процедуре МГУА исследователи предлагали свои идеи, создавая модифицированные алгоритмы для решений отдельных задач, однако их синтеза не проводилось.

МГУА относится к методам вычислительной математики [4]. При его реализации выполняется большое количество операций: решение систем линейных уравнений, вычисление псевдообратных матриц, решение задач численной оптимизации и др. Поэтому его применение без программной реализации невозможно. Любой алгоритм МГУА должен допускать эффективную программную реализацию, в том числе с применением современных компьютерных технологий: организации параллельных вычислений, использованием доступных вычислительных ресурсов и пр.

2. Общий класс задач индуктивного моделирования

Рассмотрим класс задач индуктивного моделирования, которые так или иначе сводятся к выбору оптимальной по заданному критерию модели из множества генерируемых моделей для случая объекта с одним выходом. Пусть имеется n наблюдений за поведением объекта или явления, т.е. задана выборка исходных данных $D = (Xy)$, содержащая информацию об изменении n входных признаков $X[m \times n]$ и одного выходного $y[m \times 1]$. Индуктивный процесс решения задачи определения структуры и параметров адекватной модели состоит в использовании данных одной части выборки (обучающая выборка) для создания постепенно усложняемых моделей и селекции (отбора) наиболее адекватных из них с применением принципа внешнего дополнения, выражающегося в виде ошибки моделей с использованием данных из другой части (контрольная выборка). В общем случае задача построения адекватной модели сводится к формированию по выборке экспериментальных данных некоторого множества Ф моделей различной структуры:

$$\hat{y}_f = (X, \hat{\Theta}_f), \quad f \in \Phi$$

и поиску оптимальной модели из этого множества. Проблема множественности математических моделей, используемых для описания объекта или явления, с точки зрения общих и частных целей исследования решается как задача дискретной оптимизации по условию минимума заданного внешнего критерия селекции $CR(\cdot)$:

$$f^* = \underset{f \in \Phi}{\operatorname{argmin}}\, CR(y, f(X, \hat{\Theta}_f)),$$

где оценка параметров $\hat{\Theta}_f$ для каждой $f \in \Phi$ есть решение задачи непрерывной оптимизации [5]:

$$\hat{\Theta}_f = \underset{\Theta \in R^s}{\operatorname{argmin}}\, QR(y, X, \Theta_f),$$

где $CR(\cdot) \neq QR(\cdot)$ — функционал «качества» решения задачи параметрической идентификации модели в процессе структурной идентификации, s_f — сложность модели f.

Определение произвольного многорядного алгоритма МГУА, как и вообще любой итерационной процедуры, предполагает указание:

- матрицы для начального приближения;
- оператора перехода к следующей итерации;
- правила остановки алгоритма.

3. Классический алгоритм МГУА

С учётом этого рассмотрим классический многорядный алгоритм МГУА и его некоторые авторские модификации.

1. Матрица для начального приближения алгоритма — исходная выборка из m измерений n входных признаков и выходного признака: $D = (Xy)$, где $y[m \times 1]$, $X = (x_1, x_2, \ldots, x_n)$, $x_j[m \times 1]$, $j = (1, \ldots, n)$.
2. Общий вид оператора перехода к следующему ряду алгоритма (итерации) есть некоторый функционал $\varphi : \gamma(x_i, x_j)$, где x_i, x_j — произвольная пара решений предыдущего ряда. В фиксированном классе функций задаётся функция $\hat{y} = f(x_i, x_j)$, называемая *частным описанием*. С её помощью порождаются модели на каждом ряде алгоритма. Если достоверные знания о свойствах исследуемого объекта (явления) отсутствуют, то целесообразно использовать полиномиальные функции, так как ими можно представить любую непрерывную функцию с достаточной точностью. На практике обычно используют линейно-квадратичные функции, зависящие явно от двух переменных, например:

$$\hat{y} = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2,$$

если в качестве класса функций выбран функциональный ряд Вольтерра, также называемый полиномом Колмогорова–Габора [1]:

$$y = w_0 + \sum_{i=1}^{n}(w_i x_i) + \sum_{i,j=1}^{n}(w_{ij} x_i x_j) + \sum_{i,j,k=1}^{n}(w_{ijk} x_i x_j x_k) + \ldots .$$

Оценку неизвестных параметров моделей производят, например, методом наименьших квадратов:

$$QR(\cdot) = \sum_{i=1}^{m}(\hat{y}_i(\cdot) - y_i)^2 \to \min .$$

3. Алгоритм останавливается при выполнении критерия ОСТАНОВ: $CR_{Q+1} \geqslant CR_Q$, где CR — внешний критерий качества модели, Q — номер ряда. Внешний критерий характеризует «качество» модели, а также используется для отбора «лучших» моделей на следующий ряд алгоритма. Стоит отметить, что для использования любого внешнего критерия необходимо разбить исходную выборку D на непересекающиеся обучающую T и контрольную S выборки. Приведём примеры хорошо зарекомендовавших себя внешних критериев [1–3]:
   - *критерий регулярности* — физический смысл критерия состоит в том, что он ориентирован на выбор модели, которая будет наиболее точной на множестве объектов, которых ещё нет в выборке, но они появятся в ближайшем будущем:

$$\triangle^2(T, S) = \sum_{i=1}^{m_T}(y_i^* - y_i^T)^2 + \sum_{i=1}^{m_S}(y_i^* - y_i^S)^2 \to \min,$$

где y_i^T — значения выходной величины модели, коэффициенты которой были определены по данным обучающей выборки; y_i^S — значения выходной величины модели, коэффициенты которой были определены по данным контрольной выборки; y_i^* — значения выходной величины модели на исходных данных; m_T — количество объектов (строк) в обучающей выборке; m_S — количество объектов в контрольной выборке;

– *критерий несмещённости (минимума смещения)* позволяет выбрать модель, наименее чувствительную к изменению множества опытных точек, по которым она получена. Этот критерий требует максимального совпадения выходного параметра двух моделей, полученных на данных обучающей и контрольной выборок:

$$\eta^2_{\text{см}}(T,S) = \frac{\sum\limits_{i=1}^{m}(y_i^S - y_i^T)^2}{\sum\limits_{i=1}^{m}(y_i^*)} \to \min;$$

– *комбинированный критерий*, в частном случае, представляет собой объединение показателей оценки несмещённости $\eta^2_{\text{см}}(T,S)$ и среднеквадратичной ошибки $\triangle^2(T,S)$ обеих моделей, построенных на выборках T и S, — построение одного «обобщённого» критерия:

$$\eta^2_{\text{комб}} = (1-\mu) * \eta^2_{\text{см}}(T,S) + \mu * \triangle^2(T,S) \to \min,$$

где η — коэффициент веса (экспертная оценка), представляющий объединение двух критериев. Этот подход имеет существенное допущение о «квалифицированном» эксперте и проблему компенсирования одних показателей за счёт других.

Рассмотрим подробнее процедуры, выполняемые на первом и произвольном $(Q+1)$ рядах классического многорядного алгоритма МГУА.

Первый ряд. Из входных векторов-аргументов $x_1, \ldots, x_n$ выбираются всевозможные пары $x_i \neq x_j$ и составляются частные описания, т.е. функции вида $y_l^1 = f(x_i, x_j)$, $l = 1, 2, \ldots, C_n^2$, и методами регрессионного анализа [6, 7], например, методом наименьших квадратов на обучающей выборке находятся оценки неизвестных параметров $\hat{a}_1, \hat{a}_2, \hat{a}_3, \ldots$ По заданному внешнему критерию на контрольной выборке отбираются $q * C_n^2$, где $q \in (0, 1]$ *(свобода выбора)*, лучших моделей $\hat{y}_k^1$, $k = 1, 2, \ldots, q * C_n^2$. Выполняется ряд селекций. Выходы этих $q * C_n^2$ моделей используются в качестве аргументов для формирования моделей следующего ряда. Далее находится минимальное $CR^1_{\min}$ среди всех $q * C_n^2$ значений внешнего критерия на первом ряде.

Произвольный $Q+1$ ряд. Из векторов-аргументов $\hat{y}_k^Q$, $k = 1, 2, \ldots, p$ предыдущего Q-го ряда формируются всевозможные частные описания:

$$y_l^{Q+1} = f(z_i, z_j), \quad l = 1, 2, \ldots, C_p^2, \quad z \doteq \hat{y}^Q, \quad i \neq j.$$

На обучающей выборке находятся оценки неизвестных параметров. Затем по значению внешнего критерию отбираются S «лучших» моделей из полученных частных описаний $\hat{y_l}^{Q+1}$, $l = 1, 2, \ldots, C_p^2$, находится минимальное значение $CR^{Q+1}_{\min}$. Проверяется критерий ОСТАНОВ: $CR^{Q+1}_{\min} \geqslant CR^Q_{\min}$, при выполнении которого итерационный (многорядный) процесс останавливается, иначе – переход к следующему ряду. В случае остановки в качестве оптимальной принимается модель, соответствующая значению $CR^Q_{\min}$ на предыдущем Q-м ряде.

4. Недостатки классического алгоритма и способы их устранения

В рамках данной работы обобщены значимые научно обоснованные решения исследователей и предложены авторские модификации общего классического алгоритма МГУА для устранения выявленных недостатков.

Проблему разбиения исходных данных на обучающую и контрольную выборки исследователи предлагают разрешать разными способами в зависимости от цели исследования, а также знаний о свойствах моделируемых явлений и объектов. Например, в [1] было показано, что при использовании критерия минимума смещения разделение исходных данных целесообразно производить 50% на 50% для обучающей и контрольной выборки. Исследователями предложены соответствующие методики по разбиению объектов для стационарных и динамических процессов, а также при решении задач прогнозирования и др. Автором предлагается использовать процедуру кросс-валидации (скользящего контроля), которая заключается в следующем. Фиксируется некоторое множество разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется алгоритм МГУА: генерируется модель, определяются параметры с помощью обучающей выборки, вычисляется значение внешнего критерия с использованием данных контрольной выборки. В качестве выходной модели выбирается та модель, у которой значение внешнего критерия среди выбранного множества разбиений минимально. Возможно наложение некоторых условий на множество разбиений, например, чтобы каждый объект хотя бы один раз попал в обучающую и контрольную выборки при различных разбиениях. С использованием методики скользящего контроля можно ввести дополнительную оценку полученной модели путём вычисления среднего по всем значением внешнего критерия на всех разбиениях.

В классическом многорядном алгоритме МГУА возможно ухудшение обусловленности систем уравнений с ростом числа рядов селекции. Исследователи предлагают для разрешения данной проблемы произвести нормализацию или стандартизацию исходной выборки различными способами. Например, нормализация предполагает замену номинальных признаков так, чтобы каждый из них лежал в диапазоне от 0 до 1. Стандартизация же подразумевает такую предобработку данных, после которой каждый признак имеет нулевое среднее и единичное среднеквадратическое отклонение.

Автором предлагается использовать дисперсию и её оценку (критерий Фишера) в качестве точности полученного результата, дополнительного *внутреннего* (т.е. на всех объектах исходной выборки) критерия оценки адекватности модели (соответствие математической модели исходным данным) в различных тестах при фиксации исходных данных и изменении характеристик алгоритма, например частного описания, внешнего критерия, свободы выбора и др. Также критерий Фишера может быть использован для определения оптимальной сложности структуры модели.

Итерационные алгоритмы не гарантируют построения адекватной модели, так как они базируются на неполных процедурах иерархического усложнения моделей. Возможна потеря значимых признаков, если они были исключены в первом или последующих рядах селекции. Эту проблему целесообразно решать за счёт:

- расширения порога отбора моделей-претендентов в следующий ряд алгоритма с учётом возможностей современных персональных компьютеров и общего программного обеспечения;
- полного или направленного перебора моделей-претендентов с учётом ограничений на количество исходных (промежуточных) признаков, а также возможностей персонального компьютера, общего программного обеспечения, языка программирования и др.;
- включения начальных признаков в перебор на всех рядах селекции;
- решения задачи оптимального количества моделей-претендентов, пропускаемых в следующий ряд алгоритма;

– отбора модели на ряде Q, если её эффективность по критерию отбора улучшится на $Q+1$ ряде — реализация обратной связи.

В классическом алгоритме МГУА существует возможность включения неинформативных признаков в исходную выборку и искомую модель. Для устранения этого недостатка необходимо:

– решить задачу оптимизации сложности моделей-претендентов;
– использовать методы дисперсионного анализа после завершения процедуры поиска оптимальной модели для оценки вклада в общую дисперсию каждого параметра полученной модели.

Необходимо отметить, что при квадратичном частном описании, т.е. $y(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j + a_3 x_i^2 + a_4 x_j^2$, если усложнение моделей происходит по единому правилу, показанному выше, происходит экспоненциальный рост степени. Этого недостатка можно избежать, если:

– варьировать в разном сочетании линейные, билинейные и квадратичные частные описания в соседних рядах селекции;
– изменить оператор перехода к следующему ряду алгоритма $\varphi : \gamma(x_i, x_j)$, например, для второго ряда алгоритма: $z_{klqv} = y_{kl} + y_{qv} = a_0 + a_k x_k + a_l x_l + a_q x_q + a_v x_v$ при частном описании вида $y(x_i, x_j) = a_0 + a_1 x_i + a_2 x_j$;
– решить задачу комбинаторной оптимизации структуры частных описаний.

Проблему сходимости алгоритма МГУА можно решить за счёт модификаций критерия ОСТАНОВ, например, определить ограничения на количество рядов селекции $Q_{\max}$, останавливать работу алгоритма в тот момент, как $CR_{\min}^{Q+1} - CR_{\min}^{Q} < \epsilon$. Модификация критерия ОСТАНОВ также позволит решить задачу переусложнения выходной модели.

Повысить устойчивость классического алгоритма возможно за счёт применения теории регуляризации, а именно использовать различные методы регуляризации для каждой модели-претендента на каждом ряде селекции, например, функционалы регуляризации L_1 или L_2, путём добавления «штрафного» слагаемого вида $\sum\limits_i |a_i|$ и $\sum\limits_i (a_i)^2$ соответственно при вычислении неизвестных параметров a_i.

Представить модель в явном функциональном виде можно за счёт разработки алгоритма «обратного хода», который возвращает искомую функцию, зависящую от значимых исходных входных признаков, вида $y(X)$; $X = (x_1, \ldots, x_n)$. Из решений, отобранных во всех предыдущих рядах селекции, отбрасываются все, которые не используются для образования единственного решения из последнего ряда алгоритма.

5. Заключение

Выявлены и систематизированы ключевые недостатки классического алгоритма метода группового учёта аргументов. В рамках данной работы обобщены значимые научно обоснованные решения исследователей и предложены авторские модификации классического алгоритма МГУА для устранения выявленных недостатков. Автором разработан комплекс программ, реализующий модифицированный эффективный алгоритм МГУА с рассмотренными модификациями, а также использующий методы дисперсионного анализа, корреляционного анализа, факторного анализа, элементы регрессионного анализа и др. для выделения значимых признаков, обнаружения значимых корреляций, повышения достоверности прогноза и др. [8]. Проведённые исследования и полученные практические результаты являются актуальными и могут стать основой для разработки автоматизированной системы компьютерного моделирования объектов и явлений, интеллектуального анализа и обработки данных, получения новой качественной информации и синтеза сложных систем.

Литература

1. *Ивахненко А. Г.* Системы эвристической самоорганизации в технической кибернетике. — Киев: Техніка, 1971. — 372 с.
2. *Ивахненко А. Г.* Индуктивный метод самоорганизации моделей сложных систем. — Киев: Наукова думка, 1982. — 296 с.
3. *Ивахненко А. Г., Юрчаковский Ю. П.* Моделирование сложных систем по экспериментальным данным. — М.: Радио и связь, 1978. — 120 с.
4. *Демидович Б. П., Марон И. А.* Основы вычислительной математики. — М.: Наука, 1966. — 664 с.
5. *Моисеев Н. Н., Иванилов Ю. П., Столярова Е. М.* Методы оптимизации. — М.: Наука, 1978. — 351 с.
6. *Кобзарь А. И.* Прикладная математическая статистика. — М.: Физматлит, 2006. — 816 с.
7. *Самарский А. А., Гулин А. В.* Численные методы. — М.: Наука, 1989. — 432 с.
8. *Дьячков М. Ю.* О программной реализации усовершенствованного алгоритма метода группового учета аргументов // Международная научно-методическая конференция «Некоторые вопросы анализа, алгебры, геометрии и математического образования» Воронежского государственного педагогического университета. — Воронеж: 2015. — С. 78–79.

UDC 519.6
DOI: 10.22363/2312-9735-2017-25-4-323-330

Inductive Modeling of Objects and Phenomena by the Group Method of Data Handling: the Shortcomings and Ways of Their Elimination

M. Y. Dyachkov

Department of Nonlinear Analysis and Optimization
Peoples' Friendship University of Russia (RUDN University)
6, Miklukho-Maklaya St., Moscow, Russian Federation, 117198

Original results of a research of an efficient computing method — group method of data handling are presented. Key shortcomings on each significant procedure of a classical algorithm are revealed and systematized, and also ways of their elimination, including author's modifications are presented. In particular, the use of dispersion and an assessment of dispersion (Fischer's criterion) is proposed as an assessment of accuracy of the received result, additional "internal" criterion for evaluation of adequacy of model in various tests during the fixing of input data and changing of characteristics of an algorithm, and determining the optimal complexity of the model. To solve the convergence problem of the classical algorithm, it was proposed to use the methods of dispersion, factor and correlation analysis to eliminate non-informative features, modify the criterion for stopping the algorithm. The use of regularizing functionals is suggested to solve the problem of multicollinearity of input characteristics and increase the stability of the obtained model, etc. A complex of computer modeling programs was developed, realizing an efficient modified algorithm of GMDH with the considered modifications and also methods of a dispersion analysis, correlation analysis, component analysis, elements of the regression analysis and others. The conducted researches and the received practical results can become a basis for development with use of Machine Learning and Data Science technologies of the automatic system of computer modeling, the intellectual analysis and the data processing.

Key words and phrases: mathematical modelling, inductive modeling, group method of data handling, efficient algorithm of GMDH, adequate model, complex of programs

References

1. A. G. Ivakhnenko, Systems of Heuristic Self-Organization in Technical Cybernetics, Tehnika, Kiev, 1971, in Russian.

2. A. G. Ivakhnenko, The Inductive Method of Self-Organization Models of Complex Systems, Naukova Dumka, Kiev, 1982, in Russian.
3. A. G. Ivakhnenko, Ю. П. Юрчаковский, Simulation of Complex Systems from Experimental Data, Radio i Svyaz, Moscow, 1978, in Russian.
4. B. P. Demidovich, I. A. Maron, Basics of Computational Mathematics, Nauka, M., 1966, in Russian.
5. N. N. Moiseev, Y. P. Ivanilov, E. M. Stolyarova, Optimization Methods, Nauka, Moscow, 1978, in Russian.
6. A. I. Kobzar, Applied Mathematical Statistics, FIZMATLIT, Moscow, 2006, in Russian.
7. A. A. Samarskiy, A. V. Gulin, Numerical Methods, Nauka, Moscow, 1989, in Russian.
8. M. Y. Dyachkov, On the Software Implementation of the Modified Algorithm of Group Method of Data Handling, in: International Scientific-Methodical Conference "Some Questions of Analysis, Algebra, Geometry, and Mathematical Education" of Voronezh State Pedagogical University, Voronezh, 2015, pp. 78–79, in Russian.

Для цитирования:
Дьячков М. Ю. Индуктивное моделирование объектов и явлений методом группового учёта аргументов: недостатки и способы их устранения // Вестник Российского университета дружбы народов. Серия: Математика. Информатика. Физика. — 2017. — Т. 25, № 4. — С. 323–330. — DOI: 10.22363/2312-9735-2017-25-4-323-330.

For citation:
Dyachkov M. Y. Inductive Modeling of Objects and Phenomena by the Group Method of Data Handling: the Shortcomings and Ways of Their Elimination, RUDN Journal of Mathematics, Information Sciences and Physics 25 (4) (2017) 323–330. DOI: 10.22363/2312-9735-2017-25-4-323-330. In Russian.

Сведения об авторах:
Дьячков Михаил Юрьевич — студент кафедры нелинейного анализа и оптимизации РУДН (e-mail: mihdyachkov@gmail.com, тел.: +7 (903)5411591)

Information about the authors:
Dyachkov M. Yu. — student of Nonlinear Analysis and Optimization Department of Peoples' Friendship University of Russia (RUDN University) (e-mail: mihdyachkov@gmail.com, phone: +7 (903)5411591)