ИНДУКТИВНЫЙ МЕТОД ВЫБОРА МОДЕЛИ С МИНИМАЛЬНОЙ ОШИБКОЙ И НАИМЕНЬШИМ СМЕЩЕНИЕМ ДЛЯ РЕШЕНИЯ ИНТЕРПОЛЯЦИОННЫХ ЗАДАЧ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

А.Г.Ивахненко[1], Е.А.Савченко[1], Г.А.Ивахненко[1], А.Б.Надирадзе[2], А.О.Рогов[2]

Множество объектов, подлежащих распознаванию или обнаружению зависимости, может быть замкнуто. В этом случае для выбора оптимальной модели из заданного множества моделей-кандидатов достаточно применить внешний критерий наименьшей ошибки. В случае достаточно точных данных, выбор модели неоднозначен. Тогда для доопределения модели рекомендуется применить перебор по критерию смещения. Предложен и применен новый перекрестный критерий смещения модели, расчет которого значительно более прост, чем расчет смещения моделей с помощью разделения выборки на две статистически идентичные части.

ВВЕДЕНИЕ

Интерполяционные задачи искусственного интеллекта решаются или дедуктивным, при помощи назначения сложности структуры модели человеком – автором моделирования [1] или индуктивным путем, при помощи перебора множества моделей-кандидатов по внешним критериям. Во втором случае алгоритм можно назвать самоорганизацией модели [2-5].

Системы искусственного интеллекта интерполяционного типа, в которых модели или дискриминантные функции выбраны из множества моделей-кандидатов только по одному критерию наименьшей ошибки, пригодны только для замкнутого множества входных объектов, для которого можно указать эталоны. Такие системы относятся к группе систем, получивших название систем поиска данных (*data mining*).

Если множество эталонов непрерывно расширяется, например, при распознавании рукописных знаков, то необходимо перейти к моделям, имеющим свойство обобщения, т.е. достаточно точным на выборках, которые могут быть получены на том же объекте, но которыми автор моделирования не располагает. Такие модели, подобно законам физики, должны быть несмещенными, т.е. важным для выбора оптимальной модели становится критерий смещения. Задача выбора модели решена успешно, если найдена модель достаточно точная с достаточно малым смещением. Выбор такой модели относится к области

[1] Международный научно-учебный Центр ЮНЕСКО информационных технологий и систем НАН Украины, Украина, 03187, г. Киев, Пр.Глушкова, 40, (044)266 30 28, koleso@i.kiev.ua

[2] Московский авиационный институт (государственный технический университет), Москва, 125993, Волоколамское шоссе, д.4, А-80, ГСП-3, (095)158-4674, Andrey.Nadiradze@mtu-net.ru

извлечения новых знаний (*knowledge extraction*), т.к. при этом находятся новые знания об объекте, не указанные в исходной выборке данных.

ВЫБОР ОПТИМАЛЬНОЙ МОДЕЛИ ПО КОМБИНИРОВАННОМУ КРИТЕРИЮ ИЛИ ПОСЛЕДОВАТЕЛЬНОЕ ПРИМЕНЕНИЕ ДВУХ ОПТИМИЗАЦИЙ С ДООПРЕДЕЛЕНИЕМ МОДЕЛИ

Отсутствие смещения требует, чтобы две модели, полученные на двух идентичных выборках данных (по среднему и по дисперсии переменных), были одинаковыми по всем параметрам. Это требование можно выразить математически разными способами. Например, в пространстве коэффициентов полиномиальных моделей критерий смещения будет таким:

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_Mx_M.$$

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_Mx_M.$$

$$a_0 = b_0,\ a_1 = b_1,\ a_2 = b_2,\ \ldots,\ a_M = b_M.$$

Более удобен критерий смещения, рассчитанный в пространстве ошибок:

$$ER = RR_{A/A+B} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \rightarrow \min$$

$$BS = |RR_{A/A+B} - RR_{B/A+B}| \rightarrow \min$$

Комбинированные критерии, учитывающие как ошибку, так и смещение модели можно выразить так:

$$\rho = \sqrt{\lambda \cdot ER^2 + (1-\lambda) \cdot BS^2} \rightarrow \min,$$

где λ – коэффициент веса, задаваемый автором моделирования. Описанные выше две модели соответствуют крайним случаям $\lambda=0$; $\lambda=1$. Чтобы избежать необходимости назначения коэффициента веса можно воспользоваться тем, что выбор модели по ошибке и по смещению ортогонален, т.е. не влияет друг на друга, и применить последовательно оптимизацию сначала по одному, а затем по другому критерию. При этом принцип самоорганизации модели не нарушается. Например, можно найти небольшое множество моделей, лучших по критерию ошибки, а затем выбрать из них одну, оптимальную по критерию смещения. Благоприятным обстоятельством служит то, что при переборе моделей из-за дискретного изменения сложности моделей-кандидатов, в переборной характеристике образуется интервал неопределенности выбора модели ***LC*** – ***RC***. Можно применить перебор по критерию смещения для моделей, представляющие точки которых лежат на этом интервале. В этом случае последовательное применение двух оптимизаций можно назвать доопределением оптимальной модели. Учет смещения модели используется в алгоритме три раза: 1) в неявной форме в комбинаторном алгоритме при определении оценок коэффициентов на выборке ***A*** и выбора оптимальной структуры модели на выборке ***B***; 2) в явной форме при доопределении из нескольких моделей самой несмещенной модели; 3) при выборе одной, оптимальной модели с наименьшим смещением.

Если выборка данных содержит непрерывные значения переменных, то ее разделение на две выборки с равными статистическими свойствами (средним и дисперсией) затруднено. В этом случае удобнее применить перекрестный критерий ошибки (Cross-Validation) и перекрестный критерий смещения (Cross-Bias Criteria). Расчет перекрестного критерия смещения служит как бы продолжением расчета перекрестного критерия ошибки модели. Его можно описать так.

Для расчета перекрестного критерия смещения исключаем строки последовательно по одной. На исключенных строках определяем квадрат ошибки, оцениваемой по смещению модели. При числе строк выборки равном N, получим столько же квадратов ошибки одной строки: er_{12}, er_{22}, …, er_{N2}. Далее рассчитываем значение квадратов ошибки для всех остальных строк, кроме одной, исключенной. Получим еще N значений квадратов средней ошибки:

$$ER_1^2, ER_2^2, ..., ER_N^2,\ \overline{ER} = \frac{1}{N}\sum_{i=1}^{N} er_i\ ,\ i = 1, 2, \ldots, \text{N}.$$

Если смещение модели равно нулю, то квадраты ошибки на одной строке должны быть равными среднему значению квадратов ошибок на всех остальных строках. Разность сумм квадратов ошибок дает возможность оценить смещение модели по формуле:

$$BS^2 = \frac{1}{N}\sum_{i=1}^{N}(er_i^2 - \overline{ER}^2) \rightarrow \min,\ i = 1, 2, \ldots, \text{N}.$$

Перекрестный критерий смещения может быть применен как для оценки смещения моделей, так и для автоматической кластеризации выборки данных на кластеры, необходимой, например, для многоальтернативного распознавания образов и классификации.

ВЫБОР ОПТИМАЛЬНОЙ МОДЕЛИ В АЛГОРИТМЕ ОБНАРУЖЕНИЯ ЗАВИСИМОСТИ КОЭФФИЦИЕНТА РАСПЫЛЕНИЯ ОТ ФИЗИЧЕСКИХ СВОЙСТВ МАТЕРИАЛА

Физический смысл зависимости коэффициента распыления от заданных аргументов, а также постановка задачи обнаружения этой зависимости поясняется в работе [6].

Выборка исходных данных, нормированных по наибольшему значению для каждой переменной, представлена в таблице 1. В ней обозначены: x_1 – коэффициент распыления материала ионами ксенона при энергии 300 эВ и нормальном падении ионов, мг/К; x_2 – массовая плотность, г/см^3; x_3 – молекулярный вес, а.е.м.; x_4 – температура сублимации, К; x_5 –теплоемкость Дж/моль/град; x_6 – энергия связи, эВ, а также модули коэффициента корреляции МСС каждой переменной с выходной величиной. Выходной будет служить переменная x_1. Результатом поиска будут все закономерности с достаточно малым смещением.

В данном примере получены четыре модели: две дедуктивным методом и две при помощи самоорганизации по комбинаторному алгоритму МГУА с доопределением по смещению. Результаты представлены в таблицах 2 и 3 и на рисунке в тексте.

Таблица 1. Исходные данные наблюдений взаимодействия ионов с поверхностью материалов.

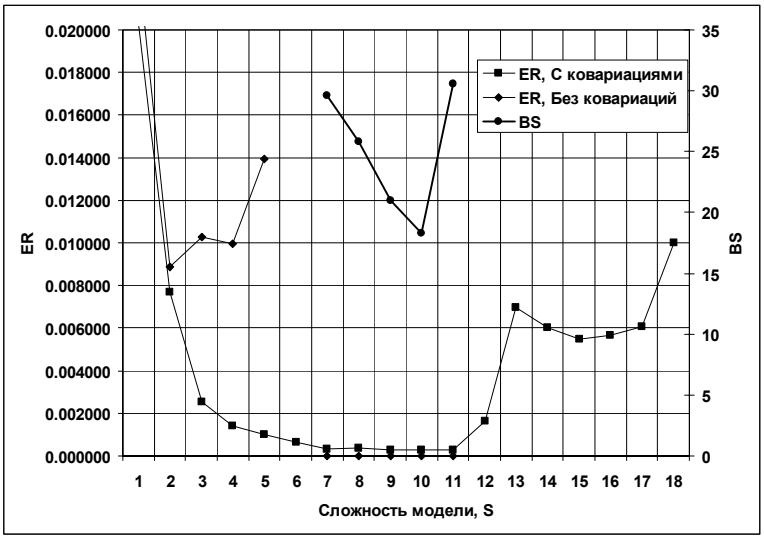
	x1	x2	x3	x4	x5	x6
M	**Y**	**Ro**	**Mw**	**Tsub**	**Cp**	**Uo**
C	0,00681100	2,300	12,0000	4473	8,536	7,410
Be	0,02281800	1,848	9,0100	2744	16,440	3,480
Si	0,11722100	2,300	28,0855	3573	19,790	3,910
Ti	0,11756600	4,505	47,8800	3560	25,060	4,340
V	0,14095600	5,960	50,9400	3665	24,480	3,700
Cr	0,23101900	7,200	51,9600	2945	23,550	3,680
Nb	0,23225000	8,570	92,9000	5073	24,440	7,500
Al	0,24074400	2,689	26,9815	2793	24,350	3,260
Fe	0,31687600	7,870	55,8470	3145	24,980	4,150
Co	0,37866000	8,900	58,9332	3230	24,800	4,380
Mo	0,38376000	10,220	95,9400	4700	23,930	6,900
Ge	0,43554000	5,323	72,5900	3120	23,220	3,770
Cu	0,65659500	8,960	63,5460	2816	24,430	3,560
Ta	1,15879700	16,650	180,9400	5623	25,290	8,700
Pd	1,17062000	12,200	106,4200	3273	25,860	4,800
Mn	1,27194500	7,320	54,9380	2353	26,280	3,150
W	1,34088800	19,340	183,8500	5953	24,270	8,760
Pt	1,65818000	21,450	195,0800	4100	25,860	5,560
Ag	1,94148000	10,500	107,8600	2440	25,360	2,700
Au	2,36352000	19,320	196,9600	3150	25,400	3,920
Max	*2,36352000*	*21,450*	*196,9600*	*5953*	*26,280*	*8,760*
MCC	1,000	0,810	0,813	0,005	0,464	0,021

Таблица 2. Выбор модели при помощи назначения предельного значения коэффициента корреляции МСС в задаче обнаружения зависимостей (без применения МГУА).

Аргументы	Оптимальная модель	ER	BS
x_i, i = 2,3,4,5,6.	$y = 0,779 – 0,220x_2 + 1,333x_3 –1,719x_4 - 0,255x_5 +0,570x_6$	0,014	0,814;
x_i; x_ix_j; i = 2,3,5; i,j = 2,2; 2,3; 2,4; 2,5; 2,6; 3,3; 3,4; 3,5; 3,6; 5,5.	$y = -0,324 + 9,98x_2 – 12,63x_3 + 0,63x_5 – 14,4x_2^2 + 22,38x_2x_3 – 57,83x_2x_4 + 1,15x_2x_5 + 47,65x_2x_6 – 8,59x_3^2 + 58,43x_3x_4 + 3,38x_3x_5 – 48,71x_3x_6 –0,482x_5^2$	0,0081	0,415

Таблица 3. Выбор модели по комбинаторному алгоритму МГУА с доопределением по смещению в задаче обнаружения зависимостей (с применением комбинаторного алгоритма МГУА)

Аргументы	Оптимальная модель	ER	BS
x_i, i = 2,3,4,5,6.	$y = 0,779 – 0,220x_2 + 1,3334x_3 –1,719x_4 - 0,255x_5 +0,570x_6$	0,014	0,814;
x_ix_j, i,j = 6; 2,2; 2,4; 2,5; 2,6; 3,4; 3,5; 4,4; 4,5; 4,6.	$y = 1,434 – 5,856x_6 + 0,430x_2^2 –4,435x_2x_4 –1,309x_2x_5 +7,891x_2x_6 –7,189x_3x_4 + 4,781x_3x_5 +0,723x_4^2 – 0,580x_4x_5 + 4,992x_4x_6$	0,00027	0,115.



Таким образом, индуктивный метод самоорганизации модели дал более эффективные результаты, чем дедуктивный метод. Основным свойством алгоритмов для осуществления интеллектуального компьютера служит свойство универсальности. Алгоритм позволяет решать все интерполяционные проблемы искусственного интеллекта. Алгоритмы комплексирования аналогов и расчета парных вероятностей [2] также имеют универсальное применение. Другим отличительным признаком интеллектуального алгоритма служит его неинтерактивный характер, то есть не требуется постоянное поступление информации от автора моделирование. Вмешательство человека в процесс перебора моделей-кандидатов не допускается, что и понимается под термином самоорганизации модели. Для осуществления самоорганизации используются две выборки данных на каждом шагу перебора, а не только в его конце для оценки точности. При этом критерий смещения применяется не менее двух раз, как показано выше, - сначала в неявной, а затем в явной форме.

СПИСОК ЛИТЕРАТУРЫ

1. Круг Г.М., Круг О.Ю. Математический метод классификации древней керамики // Труды института археологии АН СССР. – Москва: Наука, 1965.-с.317 –323.
2. Ivakhnenko A.G., Ivakhnenko G.A. and Mueller J.-A. Self-Organization of Optimum Physical Clustering of Data Sample for a Weakened Description and Forecasting of Fuzzy Objects // Pattern Recognition and Image Analysis, v.3, N4, 1993, pp. 415-422.
3. Ивахненко А.Г., Аксенова Т.И. и др. Определение кластеров активности на поверхности молекул в области заданного химического действия // Кибернетика и вычислительная техника, вып.118, 1998, стр.14-21.
4. Івахненко О.Г., Савченко Є.А., Івахненко Г. О. Алгоритм МГУА для вибору оптимальної моделі за зовнішним критерієм помилки з додатковим визначенням за зміщенням моделі та його застосування в комітетах і нейромережах // Праці І міжнародної конференції з індуктивного моделювання МКІМ, Львів 2002, І секція, с. 45 – 51.
5. A.G. Ivakhnenko, S.A. Petuhova at another. Objective Selection of Optimal Clustering of a Data Sample During Compensation of Non-Robust Random Interference // Journal of Automation and Information Sciences, 26 (3), 1993.
6. Проблемы прикладной физики. Распыление твердых тел ионной бомбардировкой. Физическое распыление одноэлементных твердых тел / Под ред. Р. Бериша: Пер. с англ. Под ред. В.А. Молчанова.- М.: 1984. - 336 с.