УДК 004

П. И. Васькин, О. Ю. Глухих
ЗАО «Морские компьютерные системы»

Псевдослучайный генератор эталонных информационных систем

Рассматривается проблема автоматического формирования тестов для сравнения алгоритмов оптимизации таблиц решений информационных систем. Дается анализ существующих подходов к сравнению эвристических алгоритмов решения NP-полных задач оптимизации. Предлагаются параметры метаинформации для псевдослучайной генерации эталонных согласованных информационных систем с многозначными атрибутами. Для задания информационных систем используется понятие многозначных многовыходных кубов. Определены операции над многозначными многовыходными кубами, которые используются в алгоритме псевдослучайного формирования эталонной таблицы решений. Приведен алгоритм псевдослучайной генерации таблиц решений эталонных информационных систем. Рассмотрен вопрос использования таблиц решений эталонных информационных систем для получения различных экземпляров таблиц решений информационной системы. Обсуждаются исследовательские задачи, которые можно решать с помощью псевдослучайного генератора эталонных информационных систем.

Информационные системы, таблицы решений, автоматическое формирование тестов, эталонные информационные системы, согласованные информационные системы, многозначные атрибуты, многозначные многовыходные кубы, комбинаторные оптимизационные задачи

Загрязнение атмосферы представляет одну из важнейших проблем современности. Повышенные концентрации диоксидов серы и азота, оксидов азота и углерода, бензапирена и формальдегида и т. п. в воздухе оказывают негативное влияние на экосистемы и здоровье людей. Особенно важно контролировать содержание вредных веществ в атмосфере населенных пунктов и вблизи промышленных объектов.

В современном мире необходимость решения оптимизационных комбинаторных задач растет с каждым днем. Будь то информационная система, модель машинного обучения или набор тестов для web-приложения – для достижения результата зачастую требуется оптимизация. Достижение того уровня предварительной обработки данных, на котором работа обучаемой модели или алгоритма анализа становится значительно эффективнее, является важным этапом извлечения знаний.

В работах, посвященных проблемам интеллектуального анализа данных и машинного обучения, широко используется понятие «информационная система» (ИС).

Информационная система представляет собой пару $I = (U, A)$, где U – конечное непустое множество объектов, также называемое универсумом, а A – конечное непустое множество атрибутов, таких, что каждый $a \in A$ определяет отображение $U \xrightarrow{a} V_a$, где V_a – множество значений атрибута a.

Обозначим через $S = (U, C, D)$ *таблицу решений* (ИС специального типа), в которой C – атрибуты условия, а D – атрибуты решения. Каждая строка этой таблицы определяет *правило принятия решений*. Правила принятия решений, которые имеют одинаковые условия (значения атрибутов условия), но разные решения (значения атрибутов решения), называются *несогласованными*, в противном случае – *согласованными*. Соответственно, такие таблицы, содержащие несовместимые правила принятия решений, называются *несогласованными*, иначе – *согласованными*.

В зависимости от того, какие значения могут принимать атрибуты, можно определить информационные системы:

– с символьными атрибутами;

– с символьными и числовыми атрибутами.

На этапе предварительной обработки данных информационной системы к наиболее важным относится проблема сокращения атрибутов. Это *NP*-полная проблема, поэтому неизбежно наличие множества разнообразных эвристических алгоритмов ее решения. Для сравнения алгоритмов друг с другом обычно используют репозитории реальных таблиц решений.

Например, в статьях по теме сокращения атрибутов зачастую для тестирования алгоритмов используются наборы данных из репозитория машинного обучения калифорнийского университета в Ирвайне (University of California and Irvine (UCI) Machine Learning Repository) [1]–[4] и репозитория KEEL (Knowledge Extraction based on Evolutionary Learning) [4], [5].

Репозиторий UCI – это коллекция баз данных и генераторов данных, которые используются сообществом машинного обучения для эмпирического анализа алгоритмов машинного обучения [1]. Архив был создан как ftp-архив в 1987 г. Дэвидом Аха и его коллегами-аспирантами в Калифорнийском университете в Ирвайне. С тех пор он широко используется студентами, преподавателями и исследователями по всему миру в качестве основного источника наборов данных машинного обучения. О влиянии архива можно судить по тому, что его цитировали более 1000 раз, что делает его одним из 100 самых цитируемых «документов» во всей компьютерной науке.

В свою очередь, KEEL – это программный инструмент с открытым исходным кодом на Java, который можно использовать для решения большого количества различных задач по обнаружению данных о знаниях. Разработчиками KEEL также предоставляется набор эталонов для анализа поведения методов машинного обучения. В частности, там находятся эталоны в формате KEEL для классификации (например, стандартных, многоэкземплярных или несбалансированных данных), полунаблюдаемой (semi-supervised) классификации, регрессии, временных рядов и ненаблюдаемого (unsupervised) обучения [5]. Кроме того, в репозитории хранится набор эталонных данных низкого качества.

Мы предлагаем для сравнения алгоритмов сокращения атрибутов ИС использовать псевдослучайный генератор информационных систем. Для этого вводится понятие эталонной таблицы решений – ИС специального типа $S_э = (U_э, C, D)$.

Эталонная информационная система (ЭИС) – это информационная система, используя которую можно получить множество таблиц решения любого объема. При подаче на вход ЭИС набора значений атрибутов условия мы получаем значения атрибутов решения. Таким образом, ЭИС заменяет экспертов и экспериментаторов, сопоставляющих значения атрибутов условия значениям атрибутов решения. Использование псевдослучайного генератора ЭИС позволяет получать исходные данные для сравнения алгоритмов сокращения атрибутов практически неограниченного объема.

Далее в статье предлагается и обсуждается состав метаинформации, которая используется в алгоритме псевдослучайного формирования эталонной таблицы решений.

Рассмотрим способ формального определения согласованной таблицы решений ИС с символьными атрибутами $S = (U, C, D)$. Для указания значений символьных атрибутов будем использовать целые числа. Пусть $C = \{c_1, c_2, ..., c_m\}$ – множество атрибутов условия. Каждый атрибут условия c_j принимает значения из множества $\{1, 2, ..., c_{j\max}\}$. Обозначим через $D = \{d_1, d_2, ..., d_l\}$ – множество атрибутов решения. Каждый атрибут решения d_k принимает значения из множества $\{1, 2, ..., d_{k\max}\}$. При формальном определении таблицы решений эталонной информационной системы будем использовать терминологию, принятую при работе с системами функций многозначной логики.

Для задания системы функций многозначной логики воспользуемся понятием многозначных многовыходных кубов.

Определение 1. Многозначный многовыходной куб – это кортеж $CD = \{c'_1, c'_2, ..., c'_m, d'_1, d'_2, ..., d'_l\}$, где каждая входная координата c'_j принимает значения из множества $\{1, 2, ..., c_{j\max}, x\}$, а каждая выходная координата d'_k – значения из множества $\{1, 2, ..., d_{k\max}, y, \overline{y}\}$. Входная координата называется свободной, если она имеет значение *x*, иначе она называется связанной. Выходная координата называется свободной, если она имеет значение *y*. Если выходная координата не свободная, то она может быть либо связанной (одно из значений множества $\{1, 2, ..., d_{k\max}\}$), либо за-

прещенной. Если выходная координата – запрещенная, то это означает, что соответствующий атрибут решения не определяется на основании значений атрибутов условия этого куба.

Свободные входные координаты куба интерпретируются следующим образом: в данном кортеже нет информации для определения зависимости функций от переменных, соответствующих этим координатам. В свою очередь, свободные выходные координаты интерпретируются следующим образом: соответствующие им функции могут принимать любое допустимое значение на входной части кортежа.

Определение 2. Обозначим через $C(D) \subseteq C$ множество входных переменных (атрибутов условия), от которых зависит хотя бы одна функция (атрибут решения) из множества D, а через $C(d_k) \subseteq C(D)$ – множество входных переменных, от которых зависит функция d_k.

Два параметра, которые обязательно должны быть включены в состав метаинформации, используемой при псевдослучайной генерации эталонных информационный систем, вполне очевидны. Это m – количество существенных входных переменных, и l – количество функций.

Так как $|C(d_k)| \le |C(D)|$ необходимо указать частоты (вероятности) того, что функции зависят от определенного числа переменных. Пусть m = = 10, а l = 5 и известно, что $|C(d_1)| = 7$, $|C(d_2)| = 6, |C(d_3)| = 8, |C(d_4)| = 6, |C(d_5)| = 4$. Таким образом, частоты событий, когда функция зависит от 4, 6, 7, 8 переменных, равны 0.2, 0.4, 0.2 и 0.2 соответственно. Отметим, что разыгрывая состав множеств $C(d_k)$, нужно следить, чтобы выполнялось условие $\left|\bigcup_{k=1}^{l} C(d_k)\right| = l$. Его выполнение означает, что каждая входная переменная будет существенной хотя бы для одной функции. Обозначим распределение вероятностей указанных событий через $\theta = \{\theta_1, \theta_2, ..., \theta_m\}$, где θ_j – вероятность того, что функция информационной системы зависит от j переменных.

Важной характеристикой информационной системы служит показатель степени пересечения множеств существенных переменных функций $C(d_k)$.

В качестве такого показателя будем использовать $\lambda = \dfrac{2\left(\sum_{i=1}^{l-1}\sum_{j=i+1}^{l} \dfrac{|C(d_i) \cap C(d_j)|}{|C(d_i) \cup C(d_j)|}\right)}{l\,(l-1)}$. Показатель λ принимает значения в диапазоне [0,1]. Если λ равно 0, то функции не имеют пересечений по множествам существенных переменных. Если λ равно 1, то все функции зависят от множества переменных $C(D)$.

При получении методом случайного разыгрывания множеств существенных переменных функций нужно понимать, что невозможно добиться одновременного выполнения двух условий:

1) точного соответствия частот мощности множеств существенных переменных функций заданным вероятностям θ;

2) близости значений заданного и полученного показателей λ.

Из этого исходит необходимость либо обозначить какой-либо фактор важным (например, частоты мощности множеств существенных переменных) и не следить за значением показателя λ, либо ввести критерий, учитывающий степень выполнения обоих условий, либо задать допустимые отклонения для частот и показателя λ и контролировать нахождение значений в определяемых ими диапазонах. Увеличение показателя λ достигается добавлением элементов в множества существенных переменных. В свою очередь, для уменьшения показателя λ необходимо исключать переменные из полученных множеств.

Прежде чем приступить к описанию алгоритма псевдослучайной генерации эталонных информационных систем, отметим следующие основные моменты:

1. Задача получения ЭИС разбивается на l независимых друг от друга задач (для каждого атрибута решения своя).

2. Каждая задача решается в два этапа. На первом этапе формируется входная часть многовыходных кубов, а выходная часть – фиксированная. Для первой задачи выходная часть равна $\{y, \overline{y}, ..., \overline{y}\}$, для второй – $\{\overline{y}, y, \overline{y}, ..., \overline{y}\}$, и т. д. Для последней задачи выходная часть равна $\{\overline{y}, \overline{y}, ..., \overline{y}, y\}$.

На втором этапе все свободные выходные координаты делаются связными (получают значения из соответствующего домена допустимых значений). Запрещенные координаты означают невозможность определения соответствующего атрибута решения на входной части данного многовыходного куба.

Пусть дан многозначный многовыходной куб $CD_1 = (c'_{1,1}, c'_{1,2}, ..., c'_{1,m}, d'_{1,1}, d'_{1,2}, ..., d'_{1,l}\}$. При решении первой задачи (получение многовыходных кубов для первого атрибута решения) выходная часть имеет вид $\{y, \overline{y}, ..., \overline{y}\}$ и не меняется в процессе решения. Дадим несколько определений, прежде чем представить алгоритм получения многовыходных кубов ЭИС.

Определение 3. Процесс формирования многовыходных кубов для выбранного атрибута решения начинается с многовыходного куба, у которого все входные координаты – свободные.

Определение 4. Элементарное преобразование многовыходного куба (назовем его γ-трансформацией) состоит в замене свободной координаты на связанные значения из области допустимых значений атрибута условия, выбранного для γ-трансформации. Свободная координата $c'_{1,j}$ порождает $c_{j\max}$ кубов заменой свободной переменной куба CD_1 на связанную координату $c'_{2,j}$.

Для изложения алгоритма будем использовать следующий пример. Пусть информационная система имеет 10 атрибутов условия и 10 атрибутов решения, причем

$$C(d_1) = \{c_1, c_2, c_3, c_4, c_5, c_7, c_8, c_9, c_{10}\};$$
$$C(d_2) = \{c_1, c_2, c_3, c_4, c_6, c_7\};$$
$$C(d_3) = \{c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}\};$$
$$C(d_4) = \{c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_{10}\};$$
$$C(d_5) = \{c_2, c_3, c_5, c_6, c_7, c_8, c_9, c_{10}\};$$
$$C(d_6) = \{c_1, c_2, c_5, c_6, c_7, c_8, c_9\};$$
$$C(d_7) = \{c_3, c_6, c_7, c_8\};\ C(d_8) = \{c_1, c_2, c_3, c_5\};$$
$$C(d_9) = \{c_1, c_5, c_8, c_9\};$$
$$C(d_{10}) = \{c_1, c_2, c_3, c_4, c_6, c_7, c_9\}.$$

Рассмотрим получение кубов для функции d_1, которая зависит от 9 переменных. С помощью псевдослучайного датчика выберем одну из переменных. Пусть это будет переменная c_5, которая может принимать 6 значений. Тогда имеем:

$$CD_1 = (xxxxxxxxxx\overline{yyyyyyyyyy});$$
$$\gamma(CD_1, c_5) = \{CD_2, CD_3, CD_4, CD_5, CD_6, CD_7\},$$

где

$$CD_2 = (xxxx1xxxxx\overline{yyyyyyyyyy});$$
$$CD_3 = (xxxx2xxxxx\overline{yyyyyyyyyy});$$
$$CD_4 = (xxxx3xxxxx\overline{yyyyyyyyyy});$$
$$CD_5 = (xxxx4xxxxx\overline{yyyyyyyyyy});$$
$$CD_6 = (xxxx5xxxxx\overline{yyyyyyyyyy});$$
$$CD_7 = (xxxx6xxxxx\overline{yyyyyyyyyy}).$$

Куб CD_1 в дальнейшем процессе трансформации больше участвует. Из оставшихся кубов псевдослучайным образом выбирается один (например, CD_6). Для выбранного куба осуществляется проверка на возможность включения в результирующее множество кубов первого атрибута решения. При этой проверке используется часть метаинформации псевдослучайного датчика ЭИС, связанная с вероятностями того, что входная часть куба содержит j связанных переменных. Пусть эти вероятности имеют следующие значения: $\rho(1) = 0$; $\rho(2) = 0.32$; $\rho(3) = 0.38$; $\rho(4) = 0.3$; $\rho(5) = 0$; $\rho(6) = 0$; $\rho(7) = 0$; $\rho(8) = 0$; $\rho(9) = 0$; $\rho(10) = 0$. Куб CD_6, имеющий одну связанную координату, не может быть включен в результирующее множество (вероятность кубов с одной связанной координатой во входной части равна нулю), поэтому его подвергаем трансформации.

Далее разыгрываем псевдослучайным образом вторую координату, которая должна стать связанной. Пусть это будет c_2, которая может принимать 4 значения. Тогда

$$\gamma(CD_6, c_2) = \{CD_8, CD_9, CD_{10}, CD_{11}),$$

где

$$CD_8 = (x1xx5xxxxx\overline{yyyyyyyyyy});$$
$$CD_9 = (x2xx5xxxxx\overline{yyyyyyyyyy});$$
$$CD_{10} = (x3xx5xxxxx\overline{yyyyyyyyyy});$$
$$CD_{11} = (x3xx5xxxxx\overline{yyyyyyyyyy}).$$

Куб CD_6 исключаем из списка кандидатов на трансформацию, поэтому после второго шага получим следующее множество кубов для γ-трансформации: $\{CD_2, CD_3, CD_4, CD_5, CD_7, CD_8, CD_9, CD_{10}, CD_{11}\}$.

Выберем псевдослучайным образом третий куб для γ-трансформации. Пусть это будет куб CD_9. Так как вероятность куба с двумя связанными координатами не равна нулю, то нужно разыграть событие, состоящее в установлении факта включения этого куба в результирующее множество кубов. Допустим, мы псевдослучайным образом разыграли, что куб, который можно включать в результат на данном шаге, должен иметь 3 связанные координаты. Куб CD_9 не соответствует этому, поэтому вы-

полним γ-трансформацию по координате c_8, например, которая может принимать два значения. Тогда

$$\gamma(CD_9, c_8) = \{CD_{12}, CD_{13}),$$

где

$$CD_{12} = (x2xx5xx1xx\underline{yyyyyyyyyy});$$
$$CD_{13} = (x2xx5xx2xx\underline{yyyyyyyyyy}).$$

После исключения куба CD_9 получим следующий список кубов для γ-трансформации: $\{CD_2, CD_3, CD_4, CD_5, CD_7, CD_8, CD_{10}, CD_{11}, CD_{12}, CD_{13}\}$.

Выберем для γ-трансформации четвертый куб. Пусть это будет куб CD_8. Далее установим факт включения этого куба в результирующее множество: куб, который можно включать в результат на данном шаге, должен иметь 2 связанные координаты. Куб CD_8 соответствует этому, поэтому γ-трансформацию его выполнять не надо, его следует включить в результирующее множество кубов и кандидатом на трансформацию он далее не будет.

Допустим, что на пятом шаге был выбран куб CD_{12}. Розыгрыш числа связанных координат в кубе для включения в результирующее множество дал значение 2. Куб CD_{12} имеет 3 связанные входные координаты, поэтому для него нужно выполнить обратную γ-трансформацию.

Определение 4. Обратная γ-трансформация – это сокращение множества кандидатов для γ-трансформации. Родитель куба, выбранного для обратной γ-трансформации, включается в результирующее множество кубов. Из множества кандидатов на γ-трансформацию исключается куб, инициирующий эту трансформацию, и все потомки его родителя.

Таким образом, на пятом шаге куб CD_8 будет помещен в результирующее множество кубов, а кубы CD_{12}, CD_{13} – исключены из множества кандидатов на трансформацию, которое приобретает следующий вид:

$$\{CD_2, CD_3, CD_4, CD_5, CD_7, CD_{10}, CD_{11}\}.$$

Процесс γ-трансформации продолжается до тех пор, пока присутствуют кандидаты на эту операцию. Затем переходим ко второй функции d_2, включив в пустой список кандидатов на трансформацию куб

$$CD_{i\,\max+1} = (xxxxxxxxxx\underline{yyyyyyyyyy}),$$

где i max – число кубов, рассмотренных при формировании результирующего множества для первого атрибута решения.

После формирования результирующего множества кубов для всех атрибутов решения заканчивается первый этап алгоритма.

Для того чтобы результирующее множество кубов имело частоты связанных координат, близкие к заданным в метаинформации вероятностям $\{\rho(n)\}$, $n \in [0, m]$, необходимо организовать пересчет этих вероятностей, используя частоты связанных координат в текущем результирующем множестве, объединенном с множеством кандидатов на γ-трансформацию. Пусть N – общее число кубов в объединении текущего результирующего множества с текущим множеством кандидатов на γ-трансформацию, а $\{\mu(n)\}$, $n \in [0, m]$ – числа кубов с n связанными координатами в этом объединенном множестве. Тогда для псевдослучайного розыгрыша числа связанных координат нужно сначала рассчитать частоты $\{\mu'(n)\}$, $n \in [0, m]$, где $\mu'(n) = (N+1) \times \rho(n) - \mu(n)$, если $(N+1) \times \rho(n) - \mu(n) \geq 0$, или $\mu'(n) = 0$ в противном случае. Затем можно вычислить вероятности чисел связанных координат: $\left\{ \rho'(n) = \dfrac{\mu'(n)}{\sum_{i=0}^{m} \mu'(n)} \right\}$, $n \in [0, m]$ и использовать их в псевдослучайном розыгрыше числа связанных координат куба, включаемого в результирующее множество.

Предложенный в данной статье алгоритм псевдослучайной генерации эталонных информационных систем был реализован на языке SQL реляционной СУБД MS SQL Server 2019. Работоспособность алгоритма была проверена на 2820 примерах, в которых число входных переменных менялось в диапазоне от 10 до 70, число функций – от 10 до 50, показатель λ – от 0.2 до 0.5, и использовались три закона распределения мощности множеств существенных переменных. Проверка показала, что параметры эталонных информационных систем, полученные по этому алгоритму, близки к параметрам метаинформации, подаваемой на вход псевдослучайного генератора эталонных информационных систем.

СПИСОК ЛИТЕРАТУРЫ

1. UCI Machine Learning Repository. URL: https://archive-beta.ics.uci.edu (дата обращения 22.07.21).

2. Thuy Ng. Sartra Wongthanavasu. A new approach for reduction of attributes based on stripped quotient sets // Pattern Recognition. 2020. Vol. 97. P. 106999.

3. Attribute reduction of SE-ISI concept lattices for incomplete contexts / Z. Wang, L. Wei, J. Qi, T. Qian // Soft computing. 2020. Vol. 24. P. 15143–15158.

4. Double-local rough sets for efficient data mining / G. Wang, T. Li, P. Zhang, H. Qian, H. Chen // Information Sciences. 2021. Vol. 571. P. 475–49.

5. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework / J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sanchez, F. Herrera // J. of Multiple-Valued Logic and Soft Computing. 2010. Vol. 17. P. 255–287.

P. I. Vaskin, O. Yu. Glukhikh
Marine computer systems CJSC

PSEUDORANDOM GENERATOR OF REFERENCE INFORMATION SYSTEMS

The problem of automatic generation of tests for comparison of algorithms for optimization of information systems' decision tables is considered. An analysis of existing approaches to comparing heuristic algorithms for solving NP-complete optimization problems is given. Metainformation parameters for pseudo-random generation of reference matched information systems with multivalued attributes are proposed. The notion of multi-valued multi-output cubes is used to define information systems. Operations on multi-valued multi-output cubes that are used in the algorithm of pseudorandom generation of the reference decision table are defined. The algorithm for pseudorandom generation of decision tables of reference information systems is presented. The question of using decision tables of reference information systems to obtain different instances of information system decision tables is considered. Research problems that can be solved with the help of a pseudorandom generator of reference information systems are discussed.

Information systems, decision tables, automatic test generation, reference information systems, matched information systems, multi-valued attributes, multi-valued multi-output cubes, combinatorial optimization problems

УДК 681.513.6

З. Х. Нгуен, В. Б. Второв
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Управление электромеханической системой с люфтом и упругими деформациями

Рассмотрена задача синтеза управления для следящей электромеханической системы с люфтом и упругостью в механической передаче при использовании адаптивно-модального регулятора и адаптивного наблюдателя состояния. Поскольку устранить незатухающие электромеханические колебания в следящей системе с упругостью и зазором средствами модального управления не удается, для повышения точности и качества динамики системы применен подход, предусматривающий использование адаптивного управления с эталонной моделью и сигнальной (релейной) адаптацией в подсистеме управления скоростью электродвигателя в сочетании с сигнальной адаптацией стационарного наблюдателя состояния электромеханического объекта с целью уменьшить вызванное нелинейностью объекта искажение оценок переменных, вырабатываемых наблюдателем. Такой подход обеспечил эффективное демпфирование возникавших колебаний и позволил значительно повысить точность следящей системы. Показан также практический подход к выбору значений весовых коэффициентов при компонентах вектора ошибки следования подсистемы управления скоростью за эталонной моделью, а также методика назначения коэффициентов усиления адаптивных цепей в наблюдателе состояния.

Упругая двухмассовая электромеханическая система, люфт, модальный регулятор, адаптивный регулятор, адаптивный наблюдатель состояния

Постановка задачи. На некоторые механизмы, например манипуляторы и самолетные антенны, налагаются требования максимального уменьшения их массы и габаритов. Это приводит