УДК 519.688

МЕТОД СЛУЧАЙНОЙ ГЕНЕРАЦИИ НЕДЕТЕРМИНИРОВАННЫХ КОНЕЧНЫХ АВТОМАТОВ И ПРОВЕРКА РЕПРЕЗЕНТАТИВНОСТИ СГЕНЕРИРОВАННЫХ СТРУКТУР

С. В. Пивнева, О. А. Рогова

Тольяттинский государственный университет

Поступила в редакцию 03.11.2010 г.

Аннотация: В статье рассматривается алгоритм случайной генерации недетерминированного конечного автомата, наиболее подходящего для выбранной предметной области. К сгенерированным структурам применяются конкретные характеристики данной предметной области и проверяется их репрезентативность.
Ключевые слова: Алгоритм, репрезентативность, случайная генерация, недетерминированный конечный автомат.
Annotation: In this paper, the algorithm of random generation of non-deterministic finite automata, more suitable for the chosen subject domain, is considered. To the generated structures, concrete characteristics of the given subject domain are applied, and their representation is tested.
Key words: Algorithm, representation, the random generation, non-deterministic finite automaton.

ВВЕДЕНИЕ

Построение адекватных математических моделей рассматриваемой предметной области является «первым шагом» математического моделирования, однако при этом одним из важнейших шагов. Этот шаг обычно «идейно близок» к проверке репрезентативности входных данных, но нередко соответствующие алгоритмы «для проверки адекватности» существенно сложнее.

Более того, с точки зрения авторов, «теоретический» и «практический» подходы к проверке адекватности предлагаемых математических моделей конкретным предметным областям также могут быть названы репрезентативностью. Поскольку одно из возможных определений последнего понятия (к сожалению, употребляемое достаточно редко) — это «возможность воспроизвести представление о целом по его части». При использовании такого определения понятие «репрезентативность» становиться более сложным, т.к. оно описывает процесс, а не объект (выборку).

Случайная генерация комбинаторных структур позволяет проверять алгоритмы, основанные на этой структуре, и исследовать поведение этих структур. Сгенерированные объекты адекватны тем потребностям, которые возникают в реальных задачах. Например, при описании контекстно-свободных языков с помощью конечных автоматов. В реальных задачах требуется достаточно большое количество состояний конечного автомата, поэтому необходимо генерировать недетерминированные конечные автоматы с целью применения к ним конкретных характеристик, которые впоследствии применяются, например, в некоторых алгоритмах LR-анализа.

МЕТОД СЛУЧАЙНОЙ ГЕНЕРАЦИИ НЕДЕТЕРМИНИРОВАННЫХ КОНЕЧНЫХ АВТОМАТОВ

В нашей работе рассматривается случайный метод генерации недетерминированного конечного автомата Ван Зиджла, основанный на случайных битовых потоках [1]. С помощью этого метода с равновероятными битовыми потоками можно успешно сравнивать различные представления регулярных языков. Но структуры, сгенерированные данным методом не соответствуют требованиям выбранной предметной области. Например, в автоматах, сгенерированных методом Ван Зиджла, сравнитель-

но мало циклов, не устраивает также разреженность автомата [1]. Поэтому был разработан другой алгоритм генерации недетерминированного конечного автомата, результатом работы которого являются автоматы, более соответствующие рассматриваемым характеристикам и подходящие для решения задач выбранной предметной области. Новый алгоритм разработан при помощи эвристик.

Новый метод, используемый для случайной генерации недетерминированного автомата в выбранной предметной области:

— задан алфавит $\sum = \{1,...,m\}$ и набор состояний $Q = \{1,...,n\}$,

— произведены равновероятные битовые потоки размера mn^2; они описывают функцию перехода δ; возникновение бита отличного от нуля в положении $2*((l-1)*n^2 + (i-1) + j$ обозначает существование перехода от состояния i в состояние j, помеченного l;

— строится трехмерная матрица переходов. В ячейках матрицы записываются символы, которые можно прочитать при переходе из состояния q_i в состояние q_j;

— есть единственное начальное состояние (вершина 1),

— набор финальных состояний случайно выбран, каждое состояние имеет равный шанс на то, чтобы быть финальным.

ПРОВЕРКА РЕПРЕЗЕНТАТИВНОСТИ СГЕНЕРИРОВАННЫХ СТРУКТУР

Для проверки репрезентативности сгенерированных структур применялись статистические критерии [3]. Если критерии $T_1, T_2, ..., T_n$ подтверждают, что последовательность ведет себя случайным образом, это не означает, что проверка с помощью T_{n+1}-го критерия будет успешной. Обычно к последовательности применяется около шести статистических критериев, и если она удовлетворяет этим критериям, то последовательность считается случайной.

Критерий «хи-квадрат». Проводим n независимых наблюдений, каждое наблюдение может принадлежать к одной из k категорий. Пусть p_s — вероятность того, что каждое наблюдение относится к категории s, пусть Y_s — число наблюдений, которые действительно относятся к категории s. Образуем статистику $V = \frac{1}{n}\sum_{s=1}^{k}\left(\frac{Y_s^2}{p_s}\right) - n$. Приемлемое значение статистики V можно определить по таблице, которая дает значения «χ^2 — распределения с v степенями свободы» для различных значений v. Используется строка таблицы с $v = k-1$, так как число «степеней свободы» равно $k-1$, что на единицу меньше, чем число категорий. Если V меньше 1 %-й точки или больше 99 %-й точки, эти числа отбрасываются как недостаточно случайные. Если V лежит между 1 %- и 5 %-й точками или между 95 %- и 99 %-й точками, то эти числа «подозрительны»; если V лежит между 5 %- и 10 %-й точками или 90 %- и 95 %-й точками, числа можно считать «почти подозрительными». Проверка по χ^2-критерию проводится три раза и более с разными данными. Если, по крайней мере, два из трех результатов оказываются подозрительными, то числа рассматриваются как недостаточно случайные.

Критерий равномерности. Проверяем равномерность распределения чисел. Выбирается число d. Для каждого r, $0 \le r < d$, подсчитывается число случаев, когда $Y_j = r$ для $0 \le j < n$, а затем применяется χ^2-критерий, принимая $k = d$ и вероятности $p_s = 1/d$ для каждой категории.

Критерий серий. Проверяем требование к последовательности, состоящее в том, чтобы пары последовательных чисел были равномерно распределены независимым образом. Подсчитываем число случаев, когда пара $(Y_{2j}, Y_{2j+1}) = (q, r)$ для $0 \le j < n$. Такая операция осуществляется для каждой пары целых чисел (q, r), таких, что $0 \le q, r < d$. Затем применяется χ^2-критерий к этим $k = d^2$ категориям, где $1/d^2$ — вероятность отнесения пары чисел к каждой из категорий.

Критерий интервалов. Этот критерий используется для проверки длины «интервалов» между появлением U_j на определенном отрезке. Если α и β — два действительных числа, таких, что $0 \le \alpha < \beta \le 1$, то рассмотрим длины подпоследовательностей $U_j, U_{j+1}, ..., U_{j+r}$, в которых U_{j+r} лежит между α и β, а другие U_s не лежат между этими числами. χ^2-критерий применяется при $k = t+1$ к значениям $count[0], count[1], ..., count[t]$ ($count[r]$ — число интервалов длиной r) с использованием следующий вероятностей: $p_r = p(1-p)^r$ для $0 \le r \le t-1$; $p_t = (1-p)^t$. Здесь $p = \beta - \alpha$ — вероятность того, что $\alpha \le U_j < \beta$. Значения n и t выбираются так, чтобы ожидаемое значение $count[r]$ равнялось 5 или больше.

Покер-критерий (критерий разбиений). Покер-критерий рассматривает n групп

по пять последовательных целых чисел $\{Y_{5j}, Y_{5j+1}, Y_{5j+2}, Y_{5j+3}, Y_{5j+4}\}$ для $0 \le j < n$ и проверяет, какие из следующих пяти категорий соответствуют таким пятеркам чисел:

5 значений = все разные;

4 значения = одна пара;

3 значения = две пары или три числа одного вида;

2 значения = полный набор или четыре числа одного вида;

1 значение = пять чисел одного вида.

В общем случае можно рассматривать n групп k последовательных чисел и подсчитывать число групп из k чисел с r различными числами. Затем применяется χ^2-критерий, в котором используются вероятности того, что в группе r различных чисел

$$p_r = \frac{d(d-1)...(d-r+1)}{d^k} \begin{Bmatrix} k \\ r \end{Bmatrix},$$

$$\begin{Bmatrix} r \\ k \end{Bmatrix} = \prod_{j=1}^{k} \frac{r+1-j}{j}.$$

Критерий собирания купонов. Используется последовательность $Y_0, Y_1, \ldots$ и находятся длины отрезков $Y_{j+1}, Y_{j+2}, \ldots, Y_{j+r}$, содержащие «полный набор» целых чисел от 0 до $d-1$.

Если дана последовательность целых чисел $Y_0, Y_1, \ldots$, таких, что $0 \le Y_j < d$, то подсчитываются длины n последовательных «собравших купоны» отрезков. $count[r]$ — это число отрезков длиной r для $d \le r < t$, а $count[t]$ — это число отрезков длиной $\ge t$.

После того как вычислено n длин, нужно применить χ^2-критерий к $count[d]$, $count[d+1], \ldots, count[t]$ с $k = t - d + 1$. Соответствующие вероятности равны

$$p_r = \frac{d!}{d^r} \begin{Bmatrix} r-1 \\ d-1 \end{Bmatrix}, \; d \le r < t;$$

$$p_t = 1 - \frac{d!}{d^{t-1}} \begin{Bmatrix} t-1 \\ d \end{Bmatrix}.$$

Критерий сериальной корреляции. Если задано n величин $U_0, U_1, \ldots, U_{n-1}$ и n других величин $V_0, V_1, \ldots, V_{n-1}$, то коэффициент корреляции между ними определяется следующим образом:

$$C = \frac{n\sum(U_j V_j) - \left(\sum U_j\right)\left(\sum V_j\right)}{\sqrt{\left(n\sum U_j^2 - \left(\sum U_j\right)^2\right)\left(n\sum V_j^2 - \left(\sum V_j\right)^2\right)}},$$

$$0 \le j < n$$

Коэффициент корреляции всегда лежит между -1 и $+1$. Когда он равен 0 или очень мал, значит, величины U_j и V_j независимы одна от другой (между ними нет линейной зависимости); если же значение коэффициента корреляции равно $+1$ или -1, это означает полную линейную зависимость

Если коэффициент корреляции равен 0 или очень мал, то величины U_j и V_j будут независимы; если же значение коэффициента корреляции равно $+1$ или -1, это означает полную линейную зависимость.

Описанные критерии были применены к сгенерированной последовательности чисел. Практическая реализация данных критериев показала, что сгенерированная последовательность чисел достаточно случайна, требования всех критериев были выполнены.

К сгенерированным недетерминированным конечным автоматам были применены рассматриваемые характеристики и получены следующие результаты:

1-я характеристика. Разреженность автомата — разработанный метод генерирует автоматы с уменьшенным средним числом дуг из вершины.

2-я характеристика. Вложенность циклов — уровень вложенности циклов увеличен.

3-я характеристика. Минимальная длина пути от стартовой вершины до финальной, деленная на количество вершин уменьшилась.

ОПИСАНИЕ ПРАКТИЧЕСКОЙ ЧАСТИ

Практическая часть работы выполнена на языке C++. Реализованы оба метода случайной генерации недетерминированного конечного автомата, а также применение конкретных характеристик данной предметной области. Проверена случайность генерируемой числовой последовательности.

Для случайной генерации недетерминированного конечного автомата используется генерация случайных чисел. Для генерации случайных чисел используются стандартные функции языка C rand и srand.

Функция rand стандартной библиотеки C генерирует целое число в диапазоне между 0 и RAND_MAX. Значение RAND_MAX должно быть по меньшей мере равно 32767 — максимальное положительное значение двухбайтового целого числа. Если rand действительно вырабатывает случайные целые числа, то при

каждом вызове rand результирующее число имеет равную вероятность оказаться любым целым, лежащим между 0 и RAND_MAX.

Функция rand на самом деле генерирует псевдослучайные числа. Повторный вызов rand производит последовательность чисел, которые кажутся случайными. Но та же самая последовательность повторяется при каждом повторении программ. Когда программа тщательно отлажена, она может быть использована для получения разных последовательностей случайных чисел при каждом выполнении.

Это рандомизация и реализуется она в законченном виде с помощью стандартной библиотечной функции srand. Функция srand получает целый аргумент unsigned и при каждом выполнении программы задает начальное число, которое функция rand использует для генерации последовательности квазислучайных чисел. Чтобы рандомизировать не вводя каждый раз начальное число, можно использовать оператор srand(time(NULL));

При этом для автоматического получения начального числа компьютер считывает показания своих часов. Функция time (с аргументом NULL) возвращает текущее «календарное время» в секундах. Это значение преобразуется в беззнаковое целое число и используется как начальное значение в генераторе случайных чисел.

Практическая реализация данных критериев показала, что сгенерированная последовательность чисел достаточно случайна, требования всех критериев были выполнены.

ЗАКЛЮЧЕНИЕ

В статье рассмотрена реализация метода случайной генерации недетерминированного конечного автомата и проверка случайности генерируемой последовательности чисел по шести статистическим критериям. Получены практические результаты, из которых видны преимущества использования разработанного метода случайной генерации недетерминированного конечного автомата в выбранной предметной области.

Программная реализация разработанного метода случайной генерации недетерминированного конечного автомата и проведенные вычислительные эксперименты показали, что сгенерированные структуры удовлетворяют требуемым характеристикам.

Сгенерированные недетерминированные автоматы репрезентативны, и в дальнейшем могут быть использованы для решения задач, возникающих в рассматриваемой предметной области.

СПИСОК ЛИТЕРАТУРЫ

1. *Мельников Б.* Недетерминированные конечные автоматы / Б. Мельников. — Тольятти: Изд-во ТГУ. — 2009. — 160 с.

2. *Champarnaud J.M.* Random generation models for NFA'S / J. M. Champarnaud, G. Hansel, T. Paranthoën, D. Ziadi. — Journal of Automata, Languages and Combinatorics, 9, 203 — 216, 2004.

3. *Кнут Д.Э.* Искусство программирования. Получисленные алгоритмы / Д. Э. Кнут. — 3-е изд. : пер. с англ. — М.: Издательский дом «Вильямс», 2003. Т. 2. — 832 с.

Пивнева Светлана Валентиновна — кандидат педагогических наук, доцент кафедры высшей математики и математического моделирования, ученый секретарь докторского диссертационного совета по физико-математическим наукам, Тольяттинский государственный университет. Тел. (8482) 53-91-17; e-mail: tlt.swetlana@rambler.ru.

Рогова Ольга Александровна — аспирантка кафедры прикладной математики и прикладной информатики, Тольяттинский государственный университет. e-mail: rogovaolgatlt@yandex.ru.

Pivneva S. V. — associated professor, department of High Mathematics and Mathematical Modelling, Togliatti State University. Tel. (8482) 53-91-17; e-mail: tlt.swetlana@rambler.ru.

Rogova O. A. — post-graduate student, department of Applied Mathematics and Applied Informatics, Togliatti State University. e-mail: rogovaolgatlt@yandex.ru.