УДК 004

Компьютерно-аналитические методы решения вероятностных задач, возникающих при исследовании случайных точечных структур*

А.Л. Резник, В.М. Ефимов, А.А. Соловьев, А.В. Торгов

Институт автоматики и электрометрии Сибирского отделения Российской академии наук (Новосибирск, Россия)

Computer Analytical Methods of Solving Probability Problems in Random Dot Patterns Research

A.L. Reznik, V.M. Efimov, A.A.Solovev, A.V. Torgov

Institute of Automation and Electrometry, Siberian Branch of the Russian Academy of Sciences (Novosibirsk, Russia)

Предложен оригинальный подход к решению весьма трудных и не имеющих на сегодня точного аналитического решения проблемных вероятностных задач, возникающих при считывании случайных точечных полей. Представлены схемы прямого, итеративного и комбинаторно-рекурсивного аналитического расчетов многомерных интегральных выражений, которыми описываются частные решения таких задач (эти решения в дальнейшем используются для нахождения общих замкнутых аналитических зависимостей). Огромный объем требующихся вычислений вынудил авторов полностью формализовать алгоритмы и перенести на ЭВМ всю тяжесть рутинных аналитических выкладок. Проведенные вычисления помогли установить (а впоследствии и доказать) целый ряд новых, ранее неизвестных вероятностных формул, характеризующих надежность считывания случайных точечных изображений, когда такое считывание проводится многоуровневыми интеграторами. Таким образом, удалось реализовать (что в научной практике случается чрезвычайно редко) идею, высказанную в свое время Дж. фон Нейманом: исследователь, встречаясь с трудной и не поддающейся решению проблемой, прибегает к компьютерным расчетам, которые «подсказывают» ему правильный ответ, а затем этот подсказанный ответ он строго доказывает. Еще одна важная особенность исследований состоит в том, что введено новое понятие «трехмерные обобщенные числа Каталана» и найден их явный вид, знание которого было эффективно использовано при решении задач, связанных с регистрацией и анализом случайных точечных изображений.

Ключевые слова: компьютерные аналитические вычисления, случайное точечное поле, многомерное интегрирование, трехмерные числа Каталана.

This paper proposes an original approach to solving complicated probability problems (there is no exact analytical solution) that arise in the reading of random point fields. The schemes of direct, iterative, and combinatorial recursive analytical calculation of multidimensional integral expression that describes the particular solutions of such problems (these solutions are used then to find the general closed analytic dependencies) are shown. A huge amount of required computations forced us to formalize all the algorithms and transfer routine analytical calculations to a computer. The calculations helped us to establish (and later prove) new set of previously unknown probabilistic formulas describing the reliability of reading random point images when such a reading is based on multilevel integrators. Thus, we were able to demonstrate the implementation of the idea proposed by John Von Neumann (extremely rare case in scientific practice): researcher meet difficult and unsolvable problem, use the computer to "suggest" him the right answer, then finds a rigorous proof. Another important feature of our study is that we introduced a new concept of "three-dimensional generalized Catalan numbers" and found their explicit form; this knowledge has been effectively used by us in solving problems related to the registration and analysis of random point images.

Key words: computer analytical calculations, random point field, multidimensional integration, three-dimensional Catalan numbers.

DOI 10.14258/izvasu(2015)1.1-32

* Работа была поддержана Российским фондом фундаментальных исследований (проект № 13-01-00361), Президиумом Российской академии наук (проект № 11/2012), Сибирским отделением Российской академии наук (проект сотрудничества СО РАН и НАН Беларуси № 16/2012).

Введение. Исследования по надежности считывания случайных точечных полей привели нас к следующей очень простой (в постановке) вероятностной задаче, имеющей непосредственное отношение к случайному разбиению интервала.

Пусть n точек $x_1, x_2, ..., x_n$ случайно брошены на интервал (0,1), т.е. имеется n независимых испытаний случайной величины, равномерно распределенной на интервале (0,1). Требуется определить вероятность $P_{n,k}(\varepsilon)$ события, состоящего в том, что не найдется ни одного подынтервала $\Omega_\varepsilon \subset (0,1)$ длины ε, содержащего более k точек.

Аналитическое решение этой одномерной задачи необходимо знать, в частности, при расчете вероятности безошибочного считывания точечных изображений, когда они формируются случайными пуассоновскими потоками постоянной интенсивности, а считывание осуществляется интеграторами, обладающими k пороговыми уровнями. Кажущаяся простота этой задачи обманчива, а ее аналитическое решение известно [1–2] лишь для $k = 1$:

$$P_{n,1}(\varepsilon) = \left(1-(n-1)\varepsilon\right)^n, \quad (0 \le \varepsilon \le 1/(n-1)). \qquad (1)$$

Следует заметить, что многие задачи, связанные со случайным разбиением интервала [3], просты в постановке, но их решение является серьезной научной проблемой. Один из путей достижения решения (1) заключается в представлении вероятности $P_{n,1}(\varepsilon)$ в виде повторного интеграла

$$P_{n,1}(\varepsilon) = n! \int\limits_{(n-1)\varepsilon}^{1} dx_n \left\{ \int\limits_{(n-2)\varepsilon}^{x_n-\varepsilon} dx_{n-1} \cdots \left\{ \int\limits_{2\varepsilon}^{x_4-\varepsilon} dx_3 \left\{ \int\limits_{\varepsilon}^{x_3-\varepsilon} dx_2 \left\{ \int\limits_{0}^{x_2-\varepsilon} dx_1 \right\} \right\} \right\} \right\}. \qquad (2)$$

Последовательное интегрирование соотношения (2) по переменным $x_1, x_2, ..., x_n$ приводит к равенству (1). Из приведенных соотношений следует, что процесс решения сформулированной выше вероятностной задачи для случая $k = 1$ весьма прост и компактен. К сожалению, для $k > 1$ вероятность $P_{n,k}(\varepsilon)$ не может быть сведена к единственному повторному интегралу, в результате чего алгоритмические трудности, которые должны быть преодолены при ее нахождении, возрастают настолько, что по состоянию на сегодня ее полного аналитического решения нет даже для $k = 2$. В настоящей работе описываются возможные подходы к решению обсуждаемой проблемы.

Программно-аналитический расчет вероятностных формул. Вероятность $P_{n,k}(\varepsilon)$ представляет собой многомерный интеграл

$$P_{n,k}(\varepsilon) = n! \int \ldots \int\limits_{D_{n,k}(\varepsilon)} dx_1 \ldots dx_n \qquad (3)$$

по области интегрирования $D_{n,k}(\varepsilon) \subset R^n$, описываемой системой линейных неравенств

$$\begin{cases} 0 < x_1 < x_2 < \ldots < x_{n-1} < x_n < 1, \\ x_{k+1} - x_1 > \varepsilon, \\ x_{k+2} - x_2 > \varepsilon, \\ \quad \vdots \\ x_n - x_{n-k} > \varepsilon. \end{cases} \qquad (4)$$

Интеграл (3) по области (4) может быть переписан в эквивалентной форме

$$P_{n,k}(\varepsilon) = n! \int\limits_{-\infty}^{+\infty} \ldots \int\limits_{-\infty}^{+\infty} 1[x_1]1[x_2]\ldots 1[x_n - x_{n-1}]1[1-x_n]1[x_{k+1} - x_1 - \varepsilon]1[x_{k+2} - x_2 - \varepsilon]\ldots 1[x_n - x_{n-k} - \varepsilon] dx_1 \ldots dx_n, \qquad (5)$$

где сомножителями в подынтегральном выражении выступают функции Хевисайда

$$1[z] = \begin{cases} 0, & z \le 0, \\ 1, & z > 0. \end{cases}$$

Затем n-мерный интеграл (5) с помощью циклического применения тождества

$$\left(\prod_{j=1}^{l} 1[x_r - \alpha_j]\right)\left(\prod_{i=1}^{m} 1[\beta_i - x_r]\right) = \sum_{j=1}^{l}\sum_{i=1}^{m} 1[x_r - \alpha_j]1[\beta_i - x_r]1[\beta_i - \alpha_j]\left(\prod_{\substack{q=1 \\ q\ne j}}^{l} 1[\alpha_j - \alpha_q]\right)\left(\prod_{\substack{s=1 \\ s\ne i}}^{m} 1[\beta_s - \beta_i]\right) \qquad (6)$$

преобразуется к набору повторных интегралов с уже расставленными пределами интегрирования (в тождестве (6) подразумевается, что выражения a и b не содержат переменной x_r).

Основная трудность практического использования приведенной процедуры состоит в том, что уже для $n = 4$ огромный объем вычислений, требующихся при расстановке пределов интегрирования, непосредственном интегрировании n-мерных повторных интегралов и проверке всех промежуточных систем неравенств на непротиворечивость, делает невозможным их проведение вручную. Поэтому нами был соз-

дан программный пакет, базирующийся на алгоритме (3)–(6) и осуществляющий все аналитические выкладки в автоматическом режиме [4].

В качестве еще одной альтернативы была предложена следующая процедура. По аналогии с известным решением (1), справедливым для $k=1$, мы попытались найти общее решение $P_{n,2}(\varepsilon)$ для $k=2$. В отличие от описанного выше алгоритма здесь была использована принципиально другая математическая техника, а именно: с помощью чисто комбинаторных средств был построен рекурсивный алгоритм, в котором формулы $P_{n,2}(\varepsilon)$ как функции непрерывного аргумента ε достигаются из дискретно-комбинаторной схемы посредством предельного перехода. При этом на каждом последующем вычислительном этапе использовались как полученные на предыдущих этапах новые результаты, так и классические комбинаторные соотношения [5].

И, наконец, третий программный пакет основан на многократном циклическом дифференцировании исходного интеграла (3) по параметру ε с дальнейшей реконструкцией вероятностных формул $P_{n,k}(\varepsilon)$ по значениям производных $\dfrac{d^{(j)}P(\varepsilon)}{d\varepsilon^{(j)}}$, $(j=0,1,\ldots,n)$ в нуле (т.е. при $\varepsilon=0$). Главное достоинство этого алгоритма заключается в том, что его применение позволяет заменить трудоемкие процедуры нахождения пределов интегрирования и последовательного многомерного интегрирования на элементарные операции подстановки и замены переменных. В данном случае это возможно благодаря применению очевидных равенств

$$\frac{d}{dz}1[z]=\delta(z), \qquad \int_{-\infty}^{\infty}\delta(z)F(z)dz=F(0),$$

которые были использованы при расчете интегралов вида (5) (здесь $\delta(z)$ – дельта-функция Дирака). Применение этого алгоритма имеет и определенные ограничения, поскольку он эффективно вычисляет формулы $P_{n,k}(\varepsilon)$ только в одном диапазоне изменения параметра ε, примыкающем к точке $\varepsilon=0$.

Используя три вышеупомянутые программные системы, мы рассчитали формулы $P_{n,k}(\varepsilon)$ для конкретных значений целочисленных переменных n и k $(k<n)$ вплоть до $n=14$ во всех диапазонах изменения непрерывного параметра ε. В дальнейшем эти компьютерные аналитические расчеты помогли нам сначала установить, а затем и строго доказать новые, ранее неизвестные вероятностные закономерности, относящиеся к случайному разбиению интервала.

Использование чисел Каталана для доказательства «компьютерных» формул. Анализ рассчитанных на компьютере формул $P_{n,k}(\varepsilon)$ позволил определить новые, ранее неизвестные общие аналитические зависимости. В частности, для четных значений $n=m$ и $k=2$ нам удалось установить формулу

$$P_{2m,2}(\varepsilon)=\frac{1}{m}C_{2m}^{m-1}\big(1-(m-1)\varepsilon\big)^{2m}, \qquad (7)$$

которая справедлива в диапазоне $1/m<\varepsilon<1/(m-1)$.

Коэффициенты $(1/m)C_{2m}^{m-1}$ в соотношении (7) являются классическими числами Каталана, известными еще по работам Леонарда Эйлера, интерес к которым сохраняется до наших дней (см., например, [6–8]), поскольку они лежат в основании перечислительной комбинаторики [9]. Любопытно отметить, что соотношение (7) было «подсказано» компьютером и опубликовано в качестве научной гипотезы более 30 лет назад [10], а строгое математическое доказательство этой формулы было получено относительно недавно [11–12]. Таким образом, мы реализовали на практике совет Дж. фон Неймана: если вы не можете найти прямое решение трудной научной проблемы, попытайтесь выполнить трудоемкие вспомогательные выкладки программно. Если вам повезет, то эти вспомогательные компьютерные вычисления «подскажут» вам правильный ответ, который вы впоследствии строго обоснуете.

Недавно нам удалось доказать [13], что вероятность $P_{n,k}(\varepsilon)$ для $k=2$ при нечетных значениях $n=2m+1$ в диапазоне $1/(m+1)<\varepsilon<1/m$ представляется в виде

$$\begin{aligned}P_{2m+1,2}(\varepsilon)&=C_{2m+1}^{m+1}(1-m\varepsilon)^{m+1}(1-(m-1)\varepsilon)^{m}-\\&-2C_{2m+1}^{m+2}(1-m\varepsilon)^{m+2}(1-(m-1)\varepsilon)^{m-1}+\\&+C_{2m+1}^{m+3}(1-m\varepsilon)^{m+3}(1-(m-1)\varepsilon)^{m-2}.\end{aligned} \qquad (8)$$

Оказалось, что найти и математически обосновать соотношение (8) много труднее, чем доказать формулу (7). Так, один из этапов этого доказательства потребовал введения нового понятия «трехмерные обобщенные числа Каталана» и определения их явной формы. В наших исследованиях эти числа возникли, когда мы столкнулись с необходимостью отыскания общего числа специальных перестановок элементов трех подмножеств, каждое из которых представляло собой ранжированную последовательность одинаково распределенных случайных величин.

Если оставить в стороне исследования, связанные со случайным разбиением интервала, то задача, которая привела нас к трехмерным числам Каталана, допускает следующую наиболее прозрачную постановку.

Необходимо найти точное число $Q_{l,m,n}$ различных слов длины $l+m+n$, которые могут быть образованы из l символов «a», m символов «b» и n символов «c» при одновременном соблюдении двух условий: 1) при просмотре слова слева направо количество встреченных символов «b» никогда не превышает количества

встреченных символов «a»; 2) при просмотре слова справа налево количество встреченных символов «c» никогда не превышает количества встреченных символов «a». Естественно, должно быть выполнено ограничение $m,\ n \le l$ (в решавшихся нами задачах со случайным разбиением интервала выполнялось более строгое условие: $m + n \le l$).

Сведя эту задачу с трехсимвольными словами к геометрической проблеме поиска путей на трехмерной дискретной решетке, нам удалось показать, что

$$Q_{l,m,n} = \frac{(l+m+n)!}{l!m!n!} - \frac{(l+m+n)!}{(l+1)!(m-1)!n!} - \frac{(l+m+n)!}{(l+1)!m!(n-1)!} + \frac{(l+m+n)!}{(l+2)!(m-1)!(n-1)!} = \\ = \frac{(l+m+n)!}{l!m!n!} \times \frac{(l+1)(l+2)-(m+n)(l+2)+mn}{(l+1)(l+2)}. \qquad (9)$$

Детальное доказательство этого равенства может быть найдено в [14]. Здесь же мы коротко опишем лишь основную суть алгоритма.

Итак, рассматриваются различные пути на трехмерной дискретной решетке в координатной системе (X, Y, Z), которые ведут из точки (0, 0, 0) в точку (l, m, n). Каждому слову ставится в соответствие один из этих путей. Символу «a» соответствует движение из текущей точки (i, j, k) в соседнюю точку (i+1, j, k); символ «b» указывает на движение в точку (i, j+1, k), а символ «c» означает движение в точку (i, j, k+1). Требуется найти общее количество таких путей из точки (0, 0, 0) в точку (l, m, n), которые не пересекают ни плоскости P_1, определяемой уравнением $X–Y=0$ (т.е. все учитываемые пути лежат в полупространстве $X \ge Y$), ни плоскости P_2, определяемой уравнением $X - Z + n - l = 0$. Таким образом, каждый из путей, удовлетворяющих обоим условиям, лежит не только в полупространстве $X \ge Y$, но также в полупространстве $l - X \ge n - Z$.

Эта переформулированная задача решается следующим образом. Из общего числа путей $S = (l+m+n)!/(l!m!n!)$, ведущих из точки (0, 0, 0) в точку (l, m, n), нужно вычесть количество путей Q, пересекающих, по крайней мере одну из плоскостей P_1 или P_2. В свою очередь для вычисления Q мы должны просуммировать количество Q_1 различных путей, пересекающих плоскость P_1, и количество Q_2 путей, пересекающих плоскость P_2, а затем отнять из результата число Q_{12} путей, пересекающих обе плоскости P_1 и P_2 (поскольку они учтены в сумме Q два раза).

Мы показали [14], что

$$Q_1 = \frac{(l+m+n)!}{(l+1)!(m-1)!n!}; \qquad Q_2 = \frac{(l+m+n)!}{(l+1)!m!(n-1)!};$$

$$Q_{12} = \frac{(l+m+n)!}{(l+2)!(m-1)!(n-1)!}.$$

Следовательно,

$$Q_{l,m,n} = S - Q_1 - Q_2 + Q_{12} = \frac{(l+m+n)!}{l!m!n!} - \frac{(l+m+n)!}{(l+1)!(m-1)!n!} - \frac{(l+m+n)!}{(l+1)!m!(n-1)!} + \\ + \frac{(l+m+n)!}{(l+2)!(m-1)!(n-1)!} = \frac{(l+m+n)!}{l!m!n!}\left[1 - \frac{m+n}{l+1} + \frac{mn}{(l+1)(l+2)}\right].$$

Таким образом, доказательство равенства (9) завершено. Напомним, что справедливость этой формулы доказывалась нами при условии $m + n \le l$, которое всегда выполнялось в рамках решавшихся нами задач, относящихся к исследованию случайных точечных полей. Если же условие $m + n \le l$ не выполняется (но, естественно, соблюдается условие $m,\ n \le l$), то решение сформулированной задачи о количестве $Q_{l,m,n}$ специальных трехсимвольных слов в общем виде запишется следующим образом:

$$Q_{l,m,n} = \frac{(l+m+n)!}{l!m!n!} - \frac{(l+m+n)!}{(l+1)!(m-1)!n!} - \frac{(l+m+n)!}{(l+1)!m!(n-1)!} + \frac{(l+m+n)!}{(l+2)!(m-1)!(n-1)!} + \\ + \frac{(l+m+n)!}{(m+n-l-2)!(l+1)!(l+1)!} - \frac{(l+m+n)!}{(m+n-l-2)!l!(l+2)!}. \qquad (10)$$

Мы назвали числа $Q_{l,m,n}$ «трехмерными обобщенными числами Каталана», имея в виду то, что эти числа расширяют традиционную последовательность Каталана, известную по многим приложениям (см., например, [15–16]) и получающуюся из равенства (10) при $n = 0$ и $l = m$. Соотношение (10) полезно не только при решении прикладных вероятностных и статистических задач, но и имеет самостоятельный теоретический интерес.

Заключение. Для нахождения точных аналитических решений проблемных вероятностных задач, возникающих при исследовании надежности считывания

случайных точечных полей, предложено и реализовано несколько программных систем, выполняющих трудоемкие аналитические выкладки. С помощью разработанного программного обеспечения был рассчитан широкий набор частных «компьютерных» решений, последующий анализ которых позволил установить (а позже и строго доказать) ряд новых, ранее неизвестных аналитических соотношений. Знание этих точных аналитических зависимостей оказывается полезным при решении многих задач, относящихся к случайному разбиению интервала.

Еще одной отличительной чертой представленной работы является то, что успешность проведенных в ней исследований в значительной мере обеспечена введением нового понятия «трехмерные обобщенные числа Каталана». Нам удалось найти простую и прозрачную интерпретацию этого естественного расширения классической последовательности Каталана, при котором обобщенные числа Каталана предстают как решение «комбинаторно-лингвистической» задачи со специальными трехсимвольными словами.

Библиографический список

1. Parzen E. Modern Probability Theory and Its Applications. John Wiley and Sons Inc. – New-York ; London, 1960.

2. Wilks S.S. Mathematical Statistics. J. Wiley and Sons. – New-York ; London, 1962.

3. David H.A., Nagaraja H.N. Order Statistics. John Wiley. – New-York, 2003.

4. Reznik A.L., Efimov V.M. Analytical Computer Calculations in Analysis of Discrete-Point Images // Pattern Recognition and Image Analysis. – 2003. – Vol. 10 (1).

5. Feller W. An introduction to probability theory and its applications. Vol. 1, 3rd ed. – New-York, 1968.

6. Stanimirovic S., Stanimirovic P., Ilic A. Ballot matrix as Catalan matrix power and related identities // Discrete Applied Mathematics. – 2012. – Vol. 160 (3).

7. Koc C., Guloglu I., Esin S. Generalized Catalan numbers, sequences and polynomials // Turk J. Math. – 2010. – Vol. 34.

8. Chamberland M., French C. Generalized Catalan Numbers and Generalized Hankel Transformations // Journal of Integer Sequences. – 2007. – Vol. 10.

9. Stanley R.P. Enumerative Combinatorics. Vol. 2. Cambridge Studies in Advanced Mathematics 62. – Cambridge, 1999.

10. Reznik A.L. Computer modeling of continuous readout of random discrete-structural images // Avtometriya. – 1981. – Vol. 6.

11. Reznik A.L., Efimov V.M., Solov'ev A.A. Computer-analytical calculation of the probability characteristics of readout of random point images // Avtometriya. – 2011. – Vol. 47 (1).

12. Reznik A.L., Efimov V.M., Torgov A.V., Solov'ev A.A. Analytical Computer Calculations in Problems with Random Division of an Interval // Pattern Recognition and Image Analysis. Advances in Mathematical Theory and Applications. – 2012. – Vol. 22 (2).

13. Reznik A.L., Efimov V.E., Solov'ev A.A., Torgov A.A. Errorless Readout of Random Discrete-Point Fields // Avtometriya. – 2012. – Vol. 48 (5).

14. Reznik A.L., Efimov V.E., Solov'ev A.A., Torgov A.V. Generalized Catalan Numbers in Problems of Processing of Random Discrete // Images Avtometriya. – 2011. – Vol. 47 (6).

15. Gardner M. Mathematical Games, Catalan numbers: an integer sequence that materializes in unexpected places // Scientific American. – 1976.

16. Hilton P., Pedersen J. Catalan numbers, their generalization, and their uses. Math. Int. – 1991. – Vol. 13.