

Программы рекурсивных аналитических вычислений в задачах анализа случайных точечных изображений*

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

Институт автоматизации и электротехники СО РАН (Новосибирск, Россия)

Programs of Recursive Analytical Calculations in Problems of Random Point Images Analysis

A.L. Reznik, A.A. Soloviev, A.V. Torgov

Institute of Automation and Electrometry SB RAS (Novosibirsk, Russia)

Обсуждается подход к решению сложных вероятностно-комбинаторных проблем, основанный на разработке и применении специализированных программных систем для проведения аналитических преобразований на ЭВМ. Рассмотрена одна из задач, связанная со случайным разбиением интервала, возникающая при исследовании надежности считывания дискретно-точечных полей и цифровых изображений, для которой удалось получить новые ранее неизвестные аналитические формулы. Эффективность разработанных программных систем обеспечивается двумя факторами: во-первых, разработкой высокоскоростных специализированных рекурсивно-комбинаторных алгоритмов; во-вторых, их программной реализацией на высокопроизводительных вычислительных кластерах с использованием современных инструментальных и языковых средств (в частности, C++ и MPI).

Приведены примеры частных решений задачи, полученных с помощью разработанных систем машинной аналитики. Продемонстрирован эффективный подход к решению сложных вероятностно-комбинаторных задач, когда компьютер выступает не просто в роли мощного вычислителя, а как эффективный помощник, обладающий обширным набором алгоритмов и программ для проведения сложных и разветвленных аналитических преобразований.

Ключевые слова: рекурсивные алгоритмы, аналитические преобразования, вычислительный кластер.

DOI 10.14258/izvasu(2020)4-18

Введение

В наших исследованиях, связанных с изучением надежности считывания случайных дискретно-точечных изображений, проводимых с помощью ин-

The paper discusses an approach to solving complex probabilistic combinatorial problems. The approach is based on the use of specialized software systems for analytical transformations for computing systems. One of the problems associated with the partition of the interval (which arises in the study of the reliability of reading discrete-point fields and digital images) is considered in the paper, and new previously unknown analytical formulas have been successfully obtained. The efficiency of the developed software systems is ensured by two factors: firstly, the development of high-speed specialized recursive-combinatorial algorithms; secondly, the software implementation on high-performance computing clusters using modern programming and development tools (such as C++ and MPI).

Examples of particular solutions to the described problem that are obtained with the help of constructed computer systems are presented. An effective approach to solving complex probabilistic combinatorial problems is demonstrated. For the proposed approach, a computer is not just a powerful "calculator", but is an effective assistant with a wide range of algorithms and programs for complex and branched analytical transformations.

Key words: recursive algorithms, analytical transformations, computing cluster.

тераторов с ограниченным числом пороговых уровней [1–2], возникла необходимость в решении одной простой по постановке, но в действительности весьма сложной вероятностно-комбинаторной задачи, об-

*Работа выполнена при поддержке РФФИ (проект №19-01-00128 А) и Министерства науки и высшего образования (проект № ААА-А17-117052410034-6).

щее решение которой на сегодняшний день неизвестно. Ее краткая формулировка такова:

«Требуется найти вероятность $P_{n,k}(\varepsilon)$ события, состоящего в том, что при случайном бросании n точек на интервал $(0,1)$ не будет образовано ни одной ε -группировки, содержащей более k точек».

Аналитическое решение этой задачи известно [3–5] лишь для случая, когда $k=1$:

$$P_{n,1}(\varepsilon) = (1 - (n-1)\varepsilon)^n, \quad (0 \leq \varepsilon \leq 1/(n-1)). \quad (1)$$

Для $k>1$ решение усложняется настолько, что в общем виде его не удастся получить даже для $k=2$. К сожалению, существующие методы решения проблем, связанных со случайным разбиением интервала [6–7], к которым относится эта задача, не позволяют вычислить вероятности $P_{n,k}(\varepsilon, L)$ при произвольных значениях n и k . Поэтому для продвижения в решении этой задачи нами было разработано несколько методов и программных систем аналитических преобразований для ЭВМ, которые позволили получить частные решения данной задачи для фиксированных параметров n и k . Их быстродействие удалось существенно повысить с применением методов параллельных вычислений. В качестве языка программирования был выбран C++ для специализированного вычислительного кластера с использованием MPI [8–10].

Аналитический расчет формул на ЭВМ

Опишем один из наиболее эффективных узкоспециализированных алгоритмов вычисления вероятностей $P_{n,k}(\varepsilon)$ для фиксированного значения $k=2$. В основе схемы лежит составление рекурсивных уравнений для дискретно-комбинаторной модели процесса регистрации случайных точечных изображений.

Итак, по аналогии с известным решением (1) для $k=1$ мы попытались найти общее аналитическое решение $P_{n,k}(\varepsilon)$ для случая $k=2$. Отличительная особенность рекурсивного подхода от других использовавшихся нами алгоритмов программно-аналитического решения поставленной задачи — при его создании использовался принципиально другой математический аппарат. На первом этапе чисто комбинаторными методами строились рекурсивные соотношения, с помощью которых в дальнейшем рассчитывались дискретные прототипы формул $P_{n,2}(\varepsilon)$. Затем для последовательно возрастающих значений целочисленного параметра n из таких дискретных аналогов-прототипов посредством предельного перехода вычислялись «непрерывные» вероятностные формулы $P_{n,2}(\varepsilon)$.

Краткая суть использованной комбинаторной модели такова. Интервал $(0,L)$ интерпретировался как совокупность r равных дискретов. Случайное бро-

сание n точек на интервал $(0,L)$ заменялось случайным бросанием n неразличимых шаров по r ящикам. Множество, объединяющее l смежных ячеек, служило аналогом подынтервала длиной ε . Бросание, в результате которого ни один из таких l -подынтервалов, содержащихся внутри исходного r -интервала $(0,L)$, не имел более 2 точек, считалось «успешным», а отношение общего числа «успешных» бросаний $Q(r,n,l)$ к общему числу исходов опыта $Q(r,n)$ принималось в качестве целочисленного аналога вероятности $P_{n,2}(\varepsilon)$. Если бы имелась возможность найти замкнутое аналитическое выражение для общего числа «успешных бросаний» $Q(r,n,l)$, то это было бы эквивалентно полному решению задачи для $k=2$. В этом случае вероятность $P_{n,2}(\varepsilon)$ вычислялась бы простым предельным переходом

$$P_{n,2}(\varepsilon) = \lim_{\substack{r \rightarrow \infty \\ (l/r) \rightarrow \varepsilon}} \frac{Q(r,n,l)}{Q(r,n)}, \quad (2)$$

поскольку выражение для общего числа бросаний $Q(r,n)$ известно [11]:

$$Q(r,n) = C_{n+r-1}^{r-1} = \frac{(n+r-1)!}{n!(r-1)!}. \quad (3)$$

Прежде чем привести рекуррентно-рекурсивные соотношения, использовавшиеся нами для расчета числа «успешных» бросаний $Q(r,n,l)$, введем ряд соглашений относительно дальнейших обозначений. Количество шаров n запишем в таком виде: $n=2N+s$; $N=0, 1, 2, \dots$; $s=1, 2$. Общее число «удачных» бросаний n шаров по r ящикам при условии, что один шар попал в 1-й ящик, один шар — в i -й, а остальные попали в ящики с номерами, большими, чем i ($i \leq l$), обозначим $q(i,r,n,l)$.

При фиксированном n функции $q(i,r,n,l)$ и $Q(r,n,l)$ представляют собой кусочно-полиномиальные выражения относительно остальных аргументов. Для полиномов, описывающих поведение этих функций на отдельных участках, введем такие обозначения:

$$\begin{aligned} q_0(i,r,n,l) &\equiv q(i,r,n,l) && \text{при } (N+2)l < r; \\ q_1(i,r,n,l) &\equiv q(i,r,n,l) && \text{при } (N+1)l < r \leq (N+2)l; \\ q_2(i,r,n,l) &\equiv q(i,r,n,l) && \text{при } Nl < r \leq (N+1)l; \\ Q_0(r,n,l) &\equiv Q(r,n,l) && \text{при } (N+1)l < r; \\ Q_1(r,n,l) &\equiv Q(r,n,l) && \text{при } Nl < r \leq (N+1)l. \end{aligned}$$

Тогда сами рекуррентные соотношения будут выглядеть следующим образом:

для $s=2$ (n четно):

$$\begin{aligned}
 q_0(i, r, n, l) &= Q_0(r-l-i, n-2, l) + \sum_{i_1=0}^i Q_0(r-2l-i_1, n-3, l) + \\
 &\quad + \sum_{i_1=0}^i \sum_{i_2=i-i_1}^l q_0(i_2, r-l-i_1, n-2, l); \\
 q_1(i, r, n, l) &= Q_0(r-l-i, n-2, l) + \sum_{i_1=0}^i \sum_{i_2=i-i_1}^{r-Nl-i_1} q_0(i_2, r-l-i_1, n-2, l); \\
 q_2(i, r, n, l) &= 0; \\
 Q_0(r, n, l) &= \sum_{k=Nl}^r \sum_{i=0}^{k-Nl} q_1(i, k, n, l); \\
 Q_1(r, n, l) &= Q_1(r-1, n, l) + Q_1(r-l, n-1, l) + \sum_{i=0}^l q_0(i, r, n, l);
 \end{aligned}
 \tag{4}$$

для $s=1$ (n нечетно):

$$\begin{aligned}
 q_0(i, r, n, l) &= Q_0(r-l-i, n-2, l) + \\
 &\quad + \sum_{i_1=0}^i Q_0(r-2l-i_1, n-3, l) + \sum_{i_1=0}^i \sum_{i_2=i-i_1}^l q_0(i_2, r-l-i_1, n-2, l); \\
 q_1(i, r, n, l) &= Q_1(r-l-i, n-2, l) + \\
 &\quad + \sum_{i_1=0}^i Q_1(r-2l-i_1, n-3, l) + \sum_{i_1=0}^i \sum_{i_2=i-i_1}^{r-Nl-i_1} q_1(i_2, r-l-i_1, n-2, l) + \\
 &\quad + \sum_{i_1=0}^i \sum_{i_2=r-Nl-i_1}^l q_2(i_2, r-l-i_1, n-2, l); \\
 q_2(i, r, n, l) &= \sum_{i_1=0}^{r-Nl} Q_1(r-2l-i_1, n-3, l) + \sum_{i_1=0}^{r-Nl} \sum_{i_2=i-i_1}^l q_2(i_2, r-l-i_1, n-2, l); \\
 Q_0(r, n, l) &= \sum_{k=Nl}^r \sum_{i=0}^{k-Nl} q_1(i, k, n, l); \\
 Q_1(r, n, l) &= \sum_{k=Nl}^r \left[Q_1(k-l, n-1, l) + \sum_{i=0}^{k-Nl} q_1(i, k, n, l) + \sum_{i=k-Nl}^l q_2(i, k, n, l) \right].
 \end{aligned}
 \tag{5}$$

По соотношениям (4) и (5) последовательно для $n=3, 4, 5, \dots$ происходит расчет формул $Q_0(r, n, l)$

и $Q_1(r, n, l)$ как функций от r и l при фиксированном значении n . Затем вычисляются пределы

$$\lim_{\substack{r \rightarrow \infty \\ l/r \rightarrow \varepsilon}} \left[\frac{Q_0(r, N, l)}{Q(r, n)} \right] \text{ и } \lim_{\substack{r \rightarrow \infty \\ l/r \rightarrow \varepsilon}} \left[\frac{Q_1(r, N, l)}{Q(r, n)} \right].
 \tag{6}$$

Начальные данные для инициации рекуррентно-рекурсивного вычислительного процесса берутся из тривиальных соображений, когда $n=0, 1, 2$. Из-за большого объема вычислений, необходимых для осуществления описанной рекурсии, весь алгоритм комбинаторного расчета был реализован в виде компьютерной программы.

Анализ «компьютерных» формул

Полученные «компьютерные» формулы позволили существенно продвинуться в решении сформулированной выше задачи. Примеры таких формул приведены в таблицах 1 и 2.

Таблица 1

Программно рассчитанные вероятностные формулы $P_{n,k}(\varepsilon)$ для $n=4$ и $k=2$

ε	$P_{4,2}(\varepsilon)$
(0, 1/2)	$1-24\varepsilon^2+12\varepsilon^3-14\varepsilon^4$
(1/2, 1)	$2-8\varepsilon+12\varepsilon^2-8\varepsilon^3+2\varepsilon^4$

Таблица 2

Программно рассчитанные вероятностные формулы $P_{n,k}(\varepsilon)$ для $n=8$ и $k=2$

ε	$P_{8,2}(\varepsilon)$
(0,1/4)	$1-168\varepsilon^2+1232\varepsilon^3-1260\varepsilon^4-18480\varepsilon^5+85064\varepsilon^6-146640\varepsilon^7+91854\varepsilon^8$
(1/4,1/3)	$14-336\varepsilon+3528\varepsilon^2-21168\varepsilon^3+79380\varepsilon^4-190512\varepsilon^5+285768\varepsilon^6-244944\varepsilon^7+91854\varepsilon^8$

Необходимо отметить, что данные таблицы 1 еще возможно рассчитать «вручную», однако уже для $n>4$ объем требуемых вычислений возрастает настолько, что без применения ЭВМ провести все необходимые выкладки невозможно.

В качестве иллюстрации к этому утверждению уместно привести следующий факт: алгоритму, использующему параллельные вычисления на современном кластере, потребовалось более 20 часов, чтобы получить аналогичную таблицу для $n=14$. Пример одного из решений приведен в таблице 3.

Таблица 3

Пример частного решения для $n=14$ и $k=2$, полученного на специализированном кластере с использованием параллельных вычислений

ε	$P_{14,2}(\varepsilon)$
(1/7,1/6)	$429-36036\varepsilon+1405404\varepsilon^2-33729696\varepsilon^3+556539984\varepsilon^4-6678479808\varepsilon^5+60106318272\varepsilon^6-412157611008\varepsilon^7+$ $+2163827457792\varepsilon^8-8655309831168\varepsilon^9+25965929493504\varepsilon^{10}-56652937076736\varepsilon^{11}-84979405615104\varepsilon^{12}-$ $-78442528260096\varepsilon^{13}+3618226397184\varepsilon^{14}$

С помощью компьютера нам удалось получить набор частных решений для всех возможных значений n и k до $n=14$ включительно. Анализ этих «компьютерных» формул позволил усмотреть общую закономерность и высказать гипотезу, заключающуюся в том, что при двух пороговых уровнях (т.е. при $k=2$) для четных n на интервале $1/(n/2) < \varepsilon < 1/((n/2) - 1)$ справедлива формула

$$P_{n,2}(\varepsilon) = (2/n)C_n^{(n/2)-1}(1 - ((n/2) - 1)\varepsilon)^n. \quad (7)$$

Дальнейшее исследование полученных на ЭВМ формул показало, что входящие в них коэффициенты составляют последовательность Каталана, что в итоге помогло в проведении строгого математического до-

казательства формулы (7) и других формул, полученных с помощью программ машинной аналитики [12].

Заключение

Отличительная особенность представленной работы в том, что компьютер в нашем случае выполнял роль не просто мощного вычислителя, но «интеллектуального» помощника, способного осуществлять расчет неизвестных вероятностных соотношений, нахождение которых невозможно выполнить вручную, так как требует сверхбольшого объема аналитических выкладок. Достигнутый прогресс позволяет надеяться, что в ближайшей перспективе нам удастся с помощью оснащения компьютера дополнительными программно-интеллектуальными вычислительными средствами решить другие сложные вероятностно-комбинаторные задачи.

Библиографический список

1. 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. DOI: 10.1134/S1054661812020125.
2. Reznik A.L., Efimov V.M., Solov'ev A.A., Torgov A.V. Errorless readout of random discrete-point fields // Optoelectronics, Instrumentation and Data Processing. 2012. Vol. 48. № 5. DOI: 10.3103/S8756699012050111.
3. Уилкс С. Математическая статистика. М., 1967.
4. Parzen E. Modern Probability Theory and Its Applications. New York, 1960. DOI: 10.1063/1.3056709.
5. Darling D.A. On Class Problems Related to the Random Division of an Interval // Annals of Mathematical Statistics. 1953. Vol. 24. DOI: 10.1214/aoms/1177729030.
6. Barton D.E., David F.N. Combinatorial Extreme Value Distributions // Mathematika. 1959. № 6. DOI: 10.1112/S0025579300001935.
7. Naus J.I. Some Probabilities, Expectations, and Variances for the Size of Largest Clusters and Smallest Intervals // The Journal of the American Statistical Association. 1966. Vol. 61. DOI: 10.1080/01621459.1966.10482203.

8. Gropp W., Lusk E., Skjellum A. Using MPI, 3rd edition: Portable Parallel Programming with the Message-Passing Interface. Cambridge, MA, 2014.

9. Оленев Н.Н. Основы параллельного программирования в системе MPI. М., 2005.

10. Pacheco P. Parallel Programming with MPI. Morgan Kaufmann, 1997.

11. Феллер В. Введение в теорию вероятностей и ее приложения. М., 1966.

12. Резник А.Л., Тузиков А.В., Соловьев А.А., Торгов А.В. Интеллектуальная программная поддержка в задачах анализа случайных цифровых изображений // Вычислительные технологии. 2018. Т. 23, № 5. DOI: 10.25743/ICT.2018.23.5.007.