©2021 г. **Ю.В. Нагребецкая**, канд. физ.-мат. наук, **В.Г. Панов**, канд. физ.-мат. наук
Е.А. Василько, Э.Э.Вахитов, Г.В. Агеев, студенты
Уральский федеральный университет, Екатеринбург,
Институт промышленной экологии УрО РАН, Екатеринбург

АЛГОРИТМЫ ВЫЧИСЛЕНИЯ СПЕКТРА ВЗАИМОДЕЙСТВИЯ ФАКТОРОВ В БИНАРНОМ ОТКЛИКЕ

Предложены алгоритмы для вычисления спектра взаимодействия бинарных факторов в бинарном отклике. Приведены оценки их временной сложности и сравнительной эффективности.
***Ключевые слова*:** булев куб, булева функция, вес Хэмминга, решетчато-упорядоченное множество, временна́я сложность алгоритма, словарь, хэш-функция, коллизия, отклик.

Y.V. Nagrebetskaya, V.G. Panov, E.A. Vasilko, E.E. Vakhitov, G.V. Ageev
Ural Federal University, Yekaterinburg,
Institute of Industrial Ecology of the Ural Branch of the Russian Academy of Sciences, Yekaterinburg

ALGORITHMS FOR CALCULATING THE SPECTRUM OF INTERACTION OF FACTORS IN BINARY RESPONSE

Algorithms for calculating the spectrum of interaction of binary factors in binary response are proposed. Estimates of their temporal complexity and comparative efficiency are given.
Key words: Boolean cube, Boolean function, Hamming weight, lattice-ordered set, time complexity of the algorithm, dictionary, hash function, collision, response.

Введение. В работах [1-5] было введено понятие спектра взаимодействия бинарных факторов в бинарном отклике, которое позволяет исчерпывающим образом описать совместное действие бинарных факторов в бинарном отклике. Рассматриваемые в этих работах математические конструкции являются элементами булевой модели бинарной теории достаточных причин, которая используется в эпидемиологии для представления причинных связей факторов, действующих на данный отклик. Ниже предложены три алгоритма для практического вычисления спектра взаимодействия.

Необходимые понятия и определения. Все необходимые обозначения и определения можно найти в [1-4], а также в тезисах [5] данного сборника.

В данной работе мы будем рассматривать различные алгоритмы вычисления спектра взаимодействия факторов в данном отклике и сравнивать их временную сложность. Входом для всех этих алгоритмов является множество C_f – носитель булевой функции *f*, а выходом - спектр M_f.

Первый алгоритм основан на определении спектра взаимодействия факторов в отклике (в булевой функции) f, зависящем от *n* факторов (булевых переменных).

Алгоритм 1.

1. Создать булев куб B^n.
2. Выделить в булевом кубе все грани текущей размерности *k* (начальное значение $k = n$).
3. Зафиксировать некоторую *k*-грань $\mathrm{B}_{\bar{I}}^{\boldsymbol{\beta}}$, где I – *k*-элементное подмножество множества $\{1,\ldots n\}$, $\boldsymbol{\beta} \in \mathrm{B}^{n-k}$.
4. Вычислить степень $\mu_{f_{I,\boldsymbol{\beta}}}$ взаимодействия *k* факторов для булевой функции $f_{I,\boldsymbol{\beta}}$ (вычисление производится по множеству $C_f \cap \mathrm{B}_{\bar{I}}^{\boldsymbol{\beta}}$).
5. Повторить пункты 3 и 4 для всех *k*-граней.

6. Выбрать максимальную степень взаимодействия *k* факторов по всем *k*-граням. Это и будет искомое число $\mu_{f,k}$.

7. Уменьшить число *k* на единицу. Если полученное значение *k* больше нуля, то перейти к пункту 2 с полученным значением *k*, иначе выдать все вычисленные значения $M_f = (\mu_{f,1},...,\mu_{f,n})$ – искомое значение спектра.

Для Алгоритма 1 произведена имплементация на языке Java.

Для ускорения работы программы все бинарные векторы-вершины кодируются их десятичными представлениями. Кроме того, используются словари, значениями которых являются списки вершин из множества $C_f \cap \mathrm{B}^{\beta}_{\bar{I}}$, а ключами – другие словари, ключами которых являются номера из множества $\bar{I}$ и значениями – соответствующие компоненты вектора $\beta \in \mathrm{B}^{n-k}$. Еще в программе используются словари для расстояний от данной вершины до каждой вершины из данного множества вершин. Существенным является то, что хэш-функции, использующиеся для всех этих словарей, не имеют коллизий.

Теорема 1. Временна́я сложность Алгоритма 1 не более, чем $O(6^n)$.

Доказательство. В п.1 алгоритма за время $O\left(n2^n\right)$ все бинарные строки-вершины булева куба B^n кодируется целыми значениями, их десятичными представлениями. Носитель C_f булевой функции (отклика) f реализуется тогда как список различных целых из множества $\{0,...,2^n-1\}$ – кодов соответствующих вершин.

Подсчитаем время работы в п. 2: (*k*-1)-грань можно получить из *k*-грани $\mathrm{B}^{\beta}_{\bar{I}}$, фиксируя переменную $x_i, i \in I$, и придавая ей значение 0 или 1. Число всех *k*-граней равно $2^{n-k}C_n^k$ [12], каждая вершина булева куба представляет собой бинарный вектор длины *n*, и в каждой *k*-грани ровно 2^k вершин. Следовательно, п. 2 выполняется за время не более, чем $O\left(2^k n2^{n-k}C_n^k\right) = O\left(n2^n C_n^k\right)$. По определению значение $\mu_{f_{I,\beta}}$ (п. 3, 4) вычисляется за время, не более, чем $O\left(4^k\right)$, а значение $\mu_{f,k}$ (п. 3-5) за время не более, чем $O\left(4^k 2^{n-k}C_n^k\right) = O\left(2^k 2^n C_n^k\right)$. Следовательно, п. 1-6 выполняются за время не более, чем

$$O\left(2^n\left(n\sum_{k=0}^{n}C_n^k + \sum_{k=0}^{n}2^k C_n^k\right)\right) \le O\left(2^n\left(n2^n + 3^n\right)\right) \le O\left(6^n\right)$$

В работе [13] приведен более эффективный **Алгоритм 2** вычисления спектра взаимодействия факторов в данном отклике и доказана следующая

Теорема 2. Временна́я сложность Алгоритма 2 не более, чем $O(n^2 4^n)$.

Алгоритм 2 распадается на два этапа [13]: (I) реализация булева куба B^n как решетчато упорядоченного множества (р.у.м.) $\left\langle \mathrm{B}^n; \le \right\rangle$; (II) непосредственное вычисление набора $M_f = (\mu_{f,1},...,\mu_{f,n})$ (т.е. спектра отклика *f*). Временна́я сложность первого этапа не более, чем $O\left(n^2 4^n\right)$, а временна́я сложность второго – не более, чем $\left|C_f\right| O\left(n^2 2^n\right)$.

В Алгоритме 3 используется более эффективная реализация первого этапа, основанная на использовании словаря и хэш-функции, не имеющей коллизий.

Алгоритм 3.

1. Создать множество B^n вершин булева куба.

2. Реализовать отношение покрытия в р.у.м. $\langle \mathrm{B}^n; \leq \rangle$.

3. Построить список $U = \{U_k\}$ ссылок на вершины, находящихся на *k*-м уровне в р.у.м. $\langle \mathrm{B}^n; \leq \rangle$, т.е. на вершины булева куба веса Хэмминга *k*.

Остальные пункты аналогичны пунктам (1)-(5) алгоритма из [13].

Для Алгоритма 3 произведена имплементация на языке C#.

Лемма. Временна́я сложность п. 1-3 Алгоритма 3 не более, чем $O(n^3 2^n)$.

Видно, что оценка временной сложности реализации р.у.м. $\langle \mathrm{B}^n; \leq \rangle$ Алгоритма 3 значительно меньше временной сложности этой реализации (I-го этапа) в Алгоритме 2: $O(n^3 2^n)$ против $O(n^2 4^n)$.

Доказательство. В п. 1 множество вершин булева куба реализовано как словарь со значениями – бинарными векторами длины *n* и ключами – соответствующими десятичными записями этих векторов. Эти десятичные целые можно считать хэш-суммами. Временна́я сложность в п.1 определения значения по ключу (перевод десятичного в двоичное) или обратно – ключа по значению (двоичное в десятичное) равна $O(n)$. В п. 2 отношение покрытия реализуются как ссылки на ключах. Просматривая каждый ключ a словаря за время $O(2^n)$, за время $O(n)$ находим его значение – бинарный вектор α. Заменяя последовательно в векторе α каждую единицу на ноль (тоже за время $O(n)$), получаем не более, чем *n* новых бинарных векторов β, покрываемых вектором α в $\langle \mathrm{B}^n; \leq \rangle$. Находим ключ b значения β (за время $O(n)$), ставим ссылку с ключа a на ключ b. Таким образом, временна́я сложность п. 2 – не более, чем $O(n^3 2^n)$. Аналогично, временная сложность п. 3 – не более, чем $O(n2^n)$. Следовательно, временна́я сложность п. 1-3 – не более, чем $O(n^3 2^n)$.

Теорема 3. На классе тех откликов, мощность носителя которых ограничена сверху линейной функцией от *n*, временна́я сложность Алгоритма 3 не более, чем $O(n^3 2^n)$.

Доказательство следует из леммы и того, что временна́я сложность этапа II Алгоритма 2 не более, чем $|C_f| O(n^2 2^n)$.

Обсуждение. Возможность реализации бинарной теории достаточных причин как приложения теории булевых алгебр и булевых функций обеспечивает возможность построения эффективных программно-вычислительных процедур для вычисления различных объектов этой теории. В частности, Алгоритм 1, основанный непосредственно на определении спектра взаимодействия бинарных факторов в данном бинарном отклике реализован на языке Java. Для ускорения работы программы использовались словари и хэш-функции, не имеющие коллизий. Описанный в [13] более эффективный Алгоритм 2 реализован на языке C# и имеет временную сложность не более,

чем $O(n^2 4^n)$. Заметим, что его оптимизация стала возможной благодаря использованию словаря и хэш-функции, не имеющей коллизий. На классе тех откликов, мощность носителя которых ограничена сверху линейной функцией от *n*, дополнительная оптимизация Алгоритма 2 позволяет значительно понизить верхнюю оценку временной сложности этого алгоритма по сравнению с ранее описанным в [13]: с $O(n^2 4^n)$ до $O(n^3 2^n)$. Полученные результаты показывают, что возможно не только формально-математическое развитие теории достаточных причин, но и реализация полученных теоретических результатов в виде эффективных алгоритмов и программ.

ЛИТЕРАТУРА

1. Нагребецкая Ю. В., Панов, В.Г. Степень взаимодействия бинарных факторов в теории достаточных причин // Системный анализ в медицине: материалы XIII междунар. конф., Благовещенск, 19-20 сентября 2019 г. Благовещенск: ДНЦ ФПД, 2019. С. 31–34.
2. Нагребецкая Ю.В., Панов В.Г. Обобщение понятия взаимодействия n факторов в теории достаточных причин и его свойства // Системный анализ в медицине: материалы XIII междунар. конф., Благовещенск, 19-20 сентября 2019 г. Благовещенск: ДНЦ ФПД, 2019. С. 35–38.
3. Nagrebetskaya J.V., Panov V.G. Joint action of binary factors in the sufficient causes theory and its classification// Int. J. Innovative Technology and Exploring Engineering. 2019. Vol. 9(1). P. 2146-2152.
4. Нагребецкая Ю.В., Панов В.Г. Максимальное взаимодействие бинарных факторов // Системный анализ в медицине: материалы XIV междунар. конф., Благовещенск, 15-16 октября 2020 г. Благовещенск: ДНЦ ФПД, 2020. С. 28–32.
5. Нагребецкая Ю.В., Панов В.Г. Число спектров взаимодействия факторов в бинарном отклике // Системный анализ в медицине: материалы XV междунар. конф., Благовещенск, 14-15 октября 2021 г. Благовещенск: ДНЦ ФПД, 2021.
6. Rothman K. Causes. Am. J. Epidemiol. 1976. Vol. 104(6). P. 587-592.
7. Miettinen O. S. Causal and preventive interdependence: Elementary principles // Scand. J. Work. Environ. Health. 1982. Vol. 8. P. 159-168.
8. Greenland S., Poole Ch. Invariants and noninvariants in the concept of interdependent effects // Scand. J. Work Environ. Health. 1988. Vol. 14. P. 125-129.
9. VanderWeele T.J., Richardson T.S. General theory for interactions in sufficient cause models with dichotomous exposures // Ann. Statistics. 2012. Vol. 40. P. 2128-2161.
10. Mackie J. L. The cement of the Universe: a study of causation. Oxford, UK: Clarendon Press, 1980.
11. Lewis D. Philosophical papers. Vol.2. Oxford, UK: Oxford University Press, 1983.
12. Гальперин Г.А. Многомерный куб // Московский центр непрерывного математического образования. 2015. С. 8-11.
13. Нагребецкая Ю.В., Панов В.Г. Оценка сложности эффективных алгоритмов проверки наличия совместного действия бинарных факторов // Системный анализ в медицине: материалы XIV междунар. конф., Благовещенск, 15-16 октября 2020 г. Благовещенск: ДНЦ ФПД, 2020. С. 36–39.

I.V.Nagrebetckaia@urfu.ru, vpanov@ecko.uran.ru,
ekaterina.vasilko@mail.ru, elnaz.v2001@gmail.com, ageev.gleb@list.ru