Е.В. Дюкова [1], *Г.О. Масляков* [2], *П.А. Прокофьев* [3]

О ПОИСКЕ МАКСИМАЛЬНЫХ НЕЗАВИСИМЫХ ЭЛЕМЕНТОВ ПРОИЗВЕДЕНИЯ ЧАСТИЧНЫХ ПОРЯДКОВ (СЛУЧАЙ ЦЕПЕЙ)*

Введение

Логический анализ данных основан на решении сложных в вычислительном плане задач, что естественно обусловлено применением дискретного аппарата. Как правило, возникают задачи, которые в теории алгоритмической сложности называют труднорешаемыми. Особой сложностью отличаются перечислительные задачи, в которых требуется найти (перечислить) все решения, при этом число решений растет экспоненциально с ростом размера задачи (размера входа). Одной из центральных перечислительных задач считается дуализация над произведением частичных порядков. Ниже приведена её формулировка.

Пусть $P = P_1 \times \ldots \times P_n$, где $P_1, \ldots, P_n$ – конечные частично упорядоченные множества. Считается, что элемент $y = \left(y_1, \ldots, y_n\right) \in P$ следует за элементом $x = \left(x_1, \ldots, x_n\right) \in P$, если y_i следует за x_i при $i = 1,2,\ldots,n$. Для обозначения того, что $y \in P$ следует за $x \in P$ и $y \neq x$ далее используется запись $x \prec y$. Пусть $R \subseteq P$, $R^+ = R \cup \{x \in P \mid \exists a \in R, a \prec x\}$. Задача построения двойственного к R множества $I\left(R\right)$, состоящего из элементов $a \in P \setminus R^+$ таких, что для любого $x \in P \setminus R^+$, $x \neq a$, отношение $a \prec x$ не выполняется, называется дуализацией над произведением частичных порядков. Элементы множества $I\left(R\right)$ называются *максимальными независимыми от R элементами P*.

*Работа поддержана грантом РФФИ № 16-01-00445

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, edjukova@mail.ru

[2]Московский государственный университет им. М.В. Ломоносова, gleb-mas@mail.ru

[3]Федеральное бюджетное учреждение науки Институт машиноведения им. А.А. Благонравова Российской академии наук, p_prok@mail.ru

Важность дуализации обусловлена большим числом приложений, среди которых прежде всего следует выделить логический анализ данных в распознавании (машинное обучение по прецедентам) и поиск ассоциативных правил в базах данных (data mining).

Одним из наиболее востребованных является случай, когда каждое P_i является цепью, т.е. любые два элемента в P_i сравнимы. Если $P_i = \{0, 1\}$ при $i \in \{1,2,\ldots,n\}$ и $0 \prec 1$, то рассматриваемая задача сводится к построению сокращённой дизъюнктивной нормальной формы монотонной булевой функции от n переменных, заданной конъюнктивной нормальной формой (КНФ) из $|R|$ элементарных дизъюнкций (дуализации монотонной КНФ). Эквивалентной задачей является поиск неприводимых покрытий булевой матрицы из $|R|$ строк и n столбцов (дуализация булевой матрицы).

Теоретические оценки эффективности алгоритмов дуализации базируются на оценке сложности одного шага [1]. Наиболее эффективным считается алгоритм, который имеет полиномиальный от размера входа шаг. На каждом шаге алгоритма строится новое решение (решение, которое не было найдено на предыдущих шагах). Такой алгоритм называется алгоритмом с полиномиальной задержкой. Однако полиномиальные алгоритмы удалось построить лишь для некоторых частных случаев дуализации монотонной КНФ. Поэтому требования к алгоритму были ослаблены. Сформировались два основных направления исследований.

Первое направление нацелено на построение так называемых инкрементальных алгоритмов, когда алгоритму разрешено просматривать решения, найденные на предыдущих шагах. При этом оценка сложности шага алгоритма даётся для худшего случая (для самого сложного варианта задачи). В [2] построен инкрементальный алгоритм дуализации монотоной КНФ с квазиполиномиальным шагом, определяемым фактически не только размером входа задачи, но и размером её выхода. В [3, 4] для случая, когда каждое P_i является цепью и $|P_i| \geq 2$, на базе алгоритма, предложенного в [2], построен квазиполиномиальный инкрементальный алгоритм. Подход интересен в основном для теории, поскольку в худшем случае число решений дуализации (размер выхода задачи) растёт экспоненциально с ростом размера её входа.

Второе направление основано на построении асимптотически оптимальных алгоритмов дуализации булевой матрицы (впервые предложено в [5]). В этом случае алгоритму разрешено делать лишние полиномиальные шаги при условии, что их число почти всегда должно

быть достаточно мало по сравнению с числом всех решений задачи (числом неприводимых покрытий булевой матрицы). В результате удалось построить алгоритмы дуализации булевой матрицы эффективные в типичном случае (эффективные для почти всех вариантов задачи). Эти алгоритмы являются лидерами по скорости счёта [6].

Теоретическое обоснование асимптотически оптимальных алгоритмов дуализации булевой матрицы базируется на получении асимптотик типичных значений числа всех неприводимых покрытий и числа шагов алгоритма. Технические основы получения подобных оценок заложены в работах [5, 7].

В [8] рассмотрен случай, когда $P_i = \{0,1,\ldots,k-1\}$, $k \geq 2$, $i = 1,2,\ldots,n$, и элементы в P_i упорядочены в порядке возрастания. Показано, что задача перечисления $I(R)$ эквивалентна построению некоторого специального подмножества множества неприводимых покрытий булевой матрицы из $|R|$ строк и kn столбцов. Для поиска элементов множества $I(R)$ построен алгоритм RUNC-M+. Данный алгоритм является модификацией асимптотически оптимального алгоритма поиска неприводимых покрытий булевой матрицы RUNC-M из [6].

В настоящей работе обоснована асимптотическая оптимальность алгоритма RUNC-M+ в случае большого числа порядков. При условии $m^{\alpha} \leq n \leq d^m$, где $\alpha > 1$, $d = k/(k-1)$, $m = |R|$, найдены асимптотики типичных значений величины $|I(R)|$ и числа шагов алгоритма. Асимптотическая оптимальность алгоритма RUNC-M+ следует из равенства полученных асимптотик.

Введено понятие упорядоченного тупикового набора столбцов целочисленной матрицы. Установлена эквивалентность задачи перечисления $I(R)$ и задачи перечисления упорядоченных тупиковых наборов столбцов матрицы L_R, строками которой являются наборы из R. В качестве иллюстративного примера дано описание асимптотически оптимального алгоритма поиска упорядоченных тупиковых наборов столбцов матрицы L_R.

1. Асимптотика типичного значения величины $\left|I\left(R\right)\right|$ в случае большого числа цепей.

Пусть $P_i = \left\{0,1,\ldots,k-1\right\}$, $k \geq 2$, $i = 1,2,\ldots,n$, и элементы в P_i упорядочены по возрастанию. Введём обозначения: L – матрица, в которой n столбцов и элементы принадлежат множеству $\left\{0,1,\ldots,k-1\right\}$, $k \geq 2$; E_k^r, $r \leq n$, – множество всех наборов вида $\left(\sigma_1,\ldots,\sigma_r\right)$, в которых $\sigma_i \in \left\{0,1,\ldots,k-1\right\}$, при $i = 1, 2,\ldots, r$.

Рассмотрим $\sigma \in E_k^r$, $\sigma = \left(\sigma_1,\ldots,\sigma_r\right)$, $\sigma_i < k-1$, $i = 1, 2,\ldots, r$. Через $Q_i(\sigma)$, $i \in \{1,2,\ldots,r\}$, обозначим множество наборов $\left(\beta_1,\ldots,\beta_r\right)$ в E_k^r таких, что $\beta_i = \sigma_i + 1$ и $\beta_j \leq \sigma_j$ при $j \in \{1,2,\ldots,r\} \setminus \{i\}$.

Пусть H – набор из r различных столбцов матрицы L. Множество различных строк подматрицы матрицы L, образованной столбцами набора H, можно рассматривать как некоторое подмножество E^H наборов из E_k^r. Набор столбцов H называется *упорядоченным тупиковым σ-покрытием* матрицы L если выполнены два следующих условия: 1) E^H не содержит набор $\left(\beta_1,\ldots,\beta_r\right) \in E_k^r$, в котором $\beta_j \leq \sigma_j$ при $j \in \{1,2,\ldots,r\}$; 2) если $i \in \{1,2,\ldots,r\}$, то E^H содержит хотя бы один набор из $Q_i(\sigma)$.

Если выполнено условие 1), то набор столбцов H называется *упорядоченным σ-покрытием* матрицы L. Если выполнено условие 2), то набор столбцов H называется *упорядоченным σ-совместимым набором столбцов* матрицы L. Упорядоченное (тупиковое) $\left(0,\ldots,0\right)$-покрытие булевой матрицы называется (*неприводимым) покрытием*.

Квадратную подматрицу порядка r матрицы L назовём *упорядоченной σ- подматрицей*, если для множества её различных строк E, рассматриваемого как некоторое подмножество наборов из E_k^r, выполнено $E \cap Q_i(\sigma) \neq \varnothing$ при $i \in \{1,2,\ldots,r\}$.

Обозначим через L_R матрицу, строками которой являются элементы множества R.

Пусть $\sigma = \left(\sigma_1,\ldots,\sigma_n\right)$ – набор из E_k^n, в котором элемент с номером t, $t \in \left\{j_1,\ldots,j_r\right\}$, не является максимальным в P_t, а элемент с

номером t, $t \notin \{j_1,\ldots,j_r\}$, является максимальным в P_t. Очевидным является

Утверждение 1. *Набор σ является максимальным независимым от R элементом множества P тогда и только тогда, когда набор столбцов матрицы L_R с номерами $j_1,\ldots,j_r$ является упорядоченным тупиковым $(\sigma_{j_1},\ldots,\sigma_{j_r})$- покрытием.*

Таким образом, задача нахождения максимальных независимых от R элементов множества P (перечисления множества $I(R)$) сводится к задаче нахождения упорядоченных тупиковых покрытий матрицы L_R.

Пусть M_{mn}^k – совокупность всех матриц размера $m \times n$ с элементами из $\{0,1,\ldots,k-1\}$, $k \geq 2$. Нас будут интересовать типичные значения числа тупиковых упорядоченных покрытий и длины тупикового упорядоченного покрытия для матрицы из M_{mn}^k. Выявление типичной ситуации будет связано с высказыванием типа «для почти всех матриц L из M_{mn}^k при $n \to \infty$ выполнено свойство P», причем свойство P может также иметь предельный характер. Например, если на матрицах из M_{mn}^k заданы две функции $F(L)$ и $G(L)$, то мы будем говорить, что для почти всех матриц L из M_{mn}^k выполнено $F(L) \sim G(L)$ ($F(L)$ асимптотически равно $G(L)$), если существуют две положительные бесконечно малые при $n \to \infty$ функции $\alpha(n)$ и $\beta(n)$ такие, что для всех достаточно больших n имеет место

$$1 - |M| / |M_{mn}^k| \leq \alpha(n),$$

где M – множество матриц L из M_{mn}^k, для которых

$$1 - \beta(n) \leq |F(L)| / |G(L)| \leq 1 + \beta(n).$$

Обозначим ϕ_d, $d > 0$, – интервал

$$\left(\frac{1}{2}\log_d mn - \frac{1}{2}\log_d \log_d mn - \log_d \log_d \log_d n,\right.$$
$$\left.\frac{1}{2}\log_d mn - \frac{1}{2}\log_d \log_d mn + \log_d \log_d \log_d n\right);$$

E_{k-1}^r – множество наборов $\left(\sigma_1,\ldots,\sigma_r\right)$ в E_k^r таких, что $\sigma_i < k-1$, $i = 1,2,\ldots,r$; $\Pi_r\left(\sigma\right) = (\sigma_1+1)^{r-1}\cdot\ldots\cdot(\sigma_r+1)^{r-1}$, $\sigma \in E_{k-1}^r$.

Пусть $L \in M_{mn}^k$, $\sigma \in E_{k-1}^r$. Положим $B\left(L,\sigma\right)$ – множество всех упорядоченных тупиковых σ-покрытий матрицы L; $S\left(L,\sigma\right)$ – множество всех упорядоченных σ-подматриц матрицы L;

$$\Sigma_1\left(L\right) = \sum_{r=1}^{n}\sum_{\sigma\in E_{k-1}^r}\left|B\left(L,\sigma\right)\right|;\ \Sigma_2\left(L\right) = \sum_{r=1}^{n}\sum_{\sigma\in E_{k-1}^r}\left|S\left(L,\sigma\right)\right|.$$

Теорема 1. *Если* $m^\alpha \le n \le d^m$, $\alpha > 1$, $d = k/(k-1)$, *то для почти всех матриц* L *из* M_{mn}^k *при* $n \to \infty$ *справедливо*

$$\Sigma_1\left(L\right) \sim \Sigma_2\left(L\right) \sim \sum_{r\in\phi_d}\sum_{\sigma\in E_{k-1}^r}\Pi_r\left(\sigma\right)C_n^rC_m^r r!\,k^{-r^2}$$

и длины почти всех упорядоченных тупиковых покрытий матрицы L *принадлежат интервалу* ϕ_d .

Замечание 1. Пусть N_{mn}^k – подмножество в M_{mn}^k, содержащее все матрицы из M_{mn}^k с попарно различными строками. В утверждении теоремы 1 можно заменить M_{mn}^k на N_{mn}^k, так как нетрудно показать, что при $m^2 = o\left(k^n\right)$, $n \to \infty$, почти все матрицы из M_{mn}^k являются матрицами с попарно различными строками.

Замечание 2. Из теоремы 1 и замечания 1 следует, что при выполнении условия теоремы 1 асимптотически оптимальным является алгоритм дуализации над произведением цепей RUNC-M+, описанный в [8] . Этот алгоритм преобразует L_R в булеву матрицу L_R^* размера из $\left|R\right| \times kn$ и перечисляет с полиномиальной задержкой некоторое специальное подмножество неприводимых покрытий матрицы L_R^*. Число шагов алгоритма RUNC-M+ не превосходит $\Sigma_2\left(L_R\right)$.

Замечание 3. Приведённые в Теореме 1 оценки для случая $k = 2$ первоначально получены в [9] с использованием понятий теории нормальных форм булевых функций. Идейно близкими являются результаты более ранней работы [5].

2. Асимптотически оптимальный алгоритм дуализации над произведением цепей

В данном разделе в качестве иллюстративного примера дано описание асимптотически оптимального алгоритма дуализации над произведением цепей *OptChains,* который в отличие от построенного в [8] алгоритма RUNC-M+ работает непосредственно с матрицей L_R (см. замечание 2 из разд. 1).

Упорядоченный $(\sigma_1,...,\sigma_r)$-совместимый набор столбцов матрицы L_R с номерами $j_1,...,j_r$ называется *максимальным*, если для любого $j_{r+1} \notin \{j_1,...,j_r\}$ и любого $\sigma_{r+1} \in P_{j_{r+1}}$ набор столбцов матрицы L_R с номерами $j_1,...,j_r,j_{r+1}$ не является упорядоченным $(\sigma_1,...,\sigma_r,\sigma_{r+1})$-совместимым.

Нетрудно видеть, что множество $G(L_R)$, состоящее из всех максимальных упорядоченных совместимых наборов столбцов матрицы L_R, содержит множество $B(L_R)$ всех упорядоченных тупиковых покрытий этой матрицы.

Алгоритм *OptChains* описывается приведённой ниже процедурой *dualOptChains*. Процедура запускается с параметрами $H = \varnothing$, $\sigma = \varnothing$, $D = \{1,...,m\}$, $C = \{1,...,n\}$, $S[j] = \{0,...,k-2\}$, $\forall j \in \{1,...,n\}$.

Процедура $dualOptChains(L_R,H,D,C,S)$

1: $J := \{j : S[j] \neq \varnothing\}$
2: **if** $J = \varnothing$ **then**
3: Сделан лишний шаг.
4: **else**
5: **for all** $j \in J$ **do**
6: $C := C \setminus \{j\}$
7: **for all** $x = k-2,...,0$ **do**
8: **if** $x \in S[j]$ **then**
9: Исключить из $S[j]$ значение x
10: $\sigma := \sigma \cup \{x\}$
11: $H := H \cup \{j\}$
12: Исключить из D номера строк i таких, что элемент матрицы L_R на пересечении i- ой строки и j- ого столбца следует за x

13: **if** $D = \varnothing$ **then**
14: Найдено упорядоченное тупиковое σ - покрытие H
15: Устранить изменения, внесённые на шагах 10, 11 и 12
16: Выйти из цикла по x
17: **else**
18: **for all** $p \in C$ **do**
19: Исключить из $S[p]$ значения y такие, что набор $H \cup \{p\}$ не является $(\sigma \cup \{y\})$- совместимым.
20: Рекурсивно вызвать $dualOptChains(L_R, H, D, C, S)$.
21: Отменить изменения, внесённые на шагах 10, 11, 12 и 19.
22: Отменить изменения, внесённые на шаге 6.

Алгоритм *OptChains* перечисляет с полиномиальной задержкой наборы столбцов из $G(L_R)$. На каждом шаге этот алгоритм за полиномиальное время строит максимальный упорядоченный σ-совместимый набор столбцов H и дополнительно за полиномиальное время осуществляет проверку, является ли H упорядоченным σ-покрытием матрицы L_R . Если H - упорядоченное σ-покрытие, то множество упорядоченных тупиковых покрытий матрицы L_R, построенное на предыдущих шагах, пополняется. В противном случае этого не происходит, так как сделан лишний шаг.

Так как $\Sigma_1(L_R) \le |G(L_R)| \le \Sigma_2(L_R)$ и $|B(L_R)| = \Sigma_1(L_R)$, то согласно теореме 1 и замечанию 1 при выполнении условия теоремы 1 число шагов алгоритма *OptChains,* равное $|G(L_R)|$, почти всегда асимптотически равно $|B(L_R)|$. Следовательно, в силу утверждения 1 алгоритм *OptChains* является асимптотически оптимальным алгоритмом дуализации над произведением цепей.

Литература

1. *Jonson D.S., Yannakakis M., Papadimitriou C.H.* On general all maximal independent sets // Information Processing Letts, 1988. Vol. 27. No. 3. P. 119–123.

2. *Fredman M., Khachiyan L.* On the complexity of dualization of monotone disjunctive normal forms // J. Algorithms, 1996. Vol. 21. P. 618–628.

3. *Boros E., Elbassioni K., Gurvich V., L. Khachiyan L., K. Makino K.* Dual-bounded generating problems: All minimal integer solutions for a

monotone system of linear inequalities // SIAM Journal on Computing, 2002. Vol. 31. No. 5. P. 1624–1643.

4. *Elbassioni, K.* Algorithms for dualization over products of partially ordered sets. *SIAM J. Discrete Math.*, 2009. Vol. 23. No. 1. P. 487–510.

5. *Дюкова Е.В.* Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР, 1977. Т. 233. № 4. С. 527–530.

6. *Дюкова Е.В., Прокофьев П.А.* Об асимптотически оптимальных алгоритмах дуализации // Журнал вычислительной математики и математической физики, 2015. Т. 55, № 5. С.895–910.

7. *Носков В.Н., Слепян В.А.* О числе тупиковых тестов для одного класса таблиц // Кибернетика, 1972. № 1. С. 60–65.

8. *Дюкова Е.В., Масляков Г.О., Прокофьев П.А.* О дуализации над произведением частичных порядков // Машинное обучение и анализ данных, 2017. Том 3. № 4. С. 239–249.

9. *Дюкова Е.В.* О сложности реализации некоторых процедур распознавания // Журнал вычислительной математики и математической физики, 1987. Том 27. № 1. С. 114–127.