О вычислительной эффективности извлечения знаний вероятностными алгоритмами

Д. В. Виноградов

Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия

Аннотация. В статье доказана вычислительная эффективность вероятностного подхода к извлечению знаний с помощью бинарной операции сходства. В дополнении к ранее доказанному автором результату о достаточности полиномиального числа гипотез о причинах исследуемого целевого свойства, в настоящей работе дана полиномиальная верхняя оценка на среднее время работы алгоритма порождения одного кандидата в гипотезы. Доказанный результат касается семейства алгоритмов, основанных на спаривающих цепях Маркова. Чтобы получить хорошую оценку на длину траектории (до попадания в эргодическое состояние) такой цепи потребовалось обогатить обучающую выборку добавлением столбцов-отрицаний для существующих бинарных признаков.

Ключевые слова: сходство, кандидат, спаривающая цепь Маркова, средняя длина траектории.

DOI 10.14357/20718594230403 **EDN** NELPQW

Введение

Извлечение знаний с использованием бинарной операции сходства началось в начале 1980-х годов в трудах В.К. Финна, предложившего ДСМ-метод автоматического порождения гипотез [1]. Этот подход получил свое имя в честь английского философа, экономиста и логика Джона Стьюарта Милля, чьи идеи по индуктивной логике [2] послужили стартовой точкой ДСМ-метода. Ключевой компонентой этого подхода является бинарная операция сходства [3]. Первоначально эта операция рассматривалась изолировано: наиболее часто – как пересечение множеств бинарных признаков, описывающих обучающие примеры. В таком случае – это был способ нахождения набора общих признаков. В дальнейшем удалось применить технику анализа формальных понятий (АФП) – раздела теории решеток [4].

Однако ДСМ-метод имеет несколько существенных ограничений, которые не позволяют ему справиться с обучающими выборками даже небольшого размера. Для преодоления этих ограничений автор [5] предложил использовать вероятностные алгоритмы. В их основе лежат операции «Замыкай по одному» CbOUp и CbODown, вторая из которых ранее встречалась в одноименном алгоритме С.О. Кузнецова [6] для исчерпывающего порождения всех кандидатов в гипотезы, число которых в некоторых случаях может оказаться экспоненциально велико. С помощью этих операций предлагается порождать малое случайное подмножество кандидатов в гипотезы о причинах исследуемого целевого свойства, каждый элемент которого соответствует одной траектории случайного блуждания по соответствующей решетке.

С использованием идей В.Н. Вапника и А. Я. Червоненкиса в работе автора [7] было показано, что достаточно породить $(n\cdot\ln 2-\ln\delta)/\varepsilon$ случайных кандидатов, чтобы с надежностью $1-\delta$ правильно доопределить ε-важные примеры.

На пути доказательства вычислительной эффективности (полиномиальной вычислимости) предложенного вероятностного метода извлечения знаний, названного ВКФ-методом, осталась единственная проблема – получение

полиномиальной границы на длину траектории алгоритма, порождающего кандидатов. В работе [5] было рассмотрено 3 алгоритма:

1. Немонотонной.
2. Монотонной.
3. Спаривающей цепи Маркова.

От первых двух пришлось отказаться, так как лежащие в их основе цепи Маркова не являются обратимыми (относительно какого-либо стационарного распределения). Получить для них полиномиальную границу на время перемешивания не удалось, что пока не позволяет оценить длину траекторий соответствующих случайных блужданий.

Третья цепь, точнее, целое семейство цепей (например, имеются также ленивый и остановленный ее варианты), имеет естественный момент остановки соответствующего алгоритма. Состояниями этой цепи (она, как и соответствующий Алгоритм 1, называется спаривающей) являются упорядоченные пары сходств, которые одновременно подвергаются преобразованию с помощью операций *CbODown* для случайно выбранного обучающего примера и *CbOUp* для случайно выбранного признака. Алгоритм спаривающей цепи Маркова останавливается тогда, когда обе компоненты состояния совпадут (склеятся).

Так как состояния с совпадающими компонентами являются эргодическими состояниями, то классическая теорема о невозвратных состояниях [8] влечет, что траектории алгоритма спаривающей цепи Маркова конечны с вероятностью 1. Таким образом, оставшимся открытым вопросом для ВКФ-метода [9] являлась задача получения полиномиальной верхней границы на среднюю длину траектории.

1. Необходимые определения и факты

Под ***обучающей выборкой*** следует понимать как бинарное отношение между элементами множества O, которые мы называем ***именами объектов*** (или просто ***объектами***), и элементами множества F, которые мы называем ***признаками***. Если в строке, соответствующей объекту $o \in O$, и столбце, соответствующем фрагменту $f \in F$, стоит единица, то мы говорим, что объект *о обладает признаком f*, и обозначаем это через $o\ I\ f$. В противном случае, говорим, что объект *о не имеет признака f*.

Для подмножества $A \subseteq O$ объектов его ***сходством*** называется подмножество $A' = \{f \in F\colon\ \forall o \in A\ [o\ I\ f]\} \subseteq F$. Поэтому $\emptyset' = F$. Для подмножества $B \subseteq F$ признаков его *сходством* называется подмножество $B' = \{o \in O\colon \forall f \in B\ [o\ I\ f]\} \subseteq O$. Теперь $\emptyset' = O$.

Определение 1. Пару $\langle A, B\rangle$ назовем ***кандидатом***, если $A = B' \subseteq O$ и $B = A' \subseteq F$.

Введем ключевое понятие для последующего изложения.

Определение 2. Операция *замыкай-по-одному-вниз* на ВКФ-кандидате $\langle A, B\rangle$ и объекте $o \in O$ порождает пару $CbODown(\langle A, B\rangle, o) = \langle (B \cap \{o\}')', B \cap \{o\}'\rangle$. Операция *замыкай-по-одному-вверх* на ВКФ-кандидате $\langle A, B\rangle$ и признаке $f \in F$ порождает пару $CbOUp(\langle A, B\rangle, f) = \langle A \cap \{f\}', (A \cap \{f\}')'\rangle$. Операция *CbODown* соответствует шагу алгоритма «Замыкай-по-одному», который был предложен С.О. Кузнецовым [6] для вычисления всех кандидатов обходом остовного дерева решетки методом «в глубину с откатом».

Определение 3. Порядок между кандидатами зададим правилом:

$\langle A, B\rangle \leq \langle C, D\rangle \Leftrightarrow B \subseteq D$.

Легко проверить, что множество всех кандидатов (для фиксированной обучающей выборки) образует решетку. Главный шаг извлечения знаний – нахождение сходств между обучающими примерами – осуществляется с помощью Алгоритма 1, основанного на спаривающей цепи Маркова.

Алгоритм 1: Спаривающая цепь Маркова.

Data: $I \subseteq O \times F$ - обучающая выборка
Result: случайный кандидат $\langle A, B\rangle$
$R := O \sqcup F$; $Min := \langle O, O'\rangle$; $Max := \langle F', F\rangle$;
while ($Min \neq Max$) **do**
    Выбираем случайный элемент $r \in R$;
    if ($r \in O$) **then**

Min := *CbODown*(*Min*, *r*); *Max* := *CbODown*(*Max*, *r*);
else
Min := *CbOUp*(*Min*, *r*); *Max* := *CbOUp*(*Max*, *r*);
end
end
⟨ *A*, *B*⟩ := *Min*;

Для доказательства корректности Алгоритма 1 нам необходима следующая лемма.

Лемма 1. *Для всякой упорядоченной пары кандидатов* ⟨ *A*, *B*⟩ ≤ ⟨ *C*, *D*⟩ *и любого o* ∈ *O имеем:*

$$CbODown(\langle A, B\rangle, o) \le CbODown(\langle C, D\rangle, o).$$

Для всякой упорядоченной пары кандидатов ⟨ *A*, *B*⟩ ≤ ⟨ *C*, *D*⟩ *и любого f* ∈ *F имеем:*

$$CbOUp(\langle A, B\rangle, f) \le CbOUp(\langle C, D\rangle, f).$$

Определение 4. Множество состояний $E = \{(\langle A, B\rangle = \langle A, B\rangle): \langle A, B\rangle\}$ спаривающей цепи Маркова (совпадающих пар кандидатов) называется ***эргодическим множеством***. Состояния $s_i \in E$ называются ***эргодическими***. Состояние вида $(\langle A, B\rangle < \langle C, D\rangle)$ называется ***невозвратным***. Классическая теорема о невозвратных состояниях цепи Маркова [8] может быть сформулирована как утверждение:

$$lim_{\,t\to\infty}\, \mathbf{P}[X_t \notin E | X_0 = s_i] \to 0 \text{ для любого } s_i \notin E. \quad (1)$$

Из нее легко выводится одна из ключевых теорем.

Теорема 1. *Алгоритм* 1 *останавливается с вероятностью* 1.

Доказательство. Мы докажем немного более общее утверждение о том, что моменты $T_i(E) = \min\{t: X_t \in E, X_0 = s_i\}$ первого попадания в E, стартуя с некоего невозвратного состояния $s_i = (\langle A, B\rangle < \langle C, D\rangle) \notin E$, являются *марковскими*, т.е. $\mathbf{P}[T_i(E) < \infty | X_0 = s_i] = 1$.

Имеем $\{X_t \in E, X_0 = s_i\} = \cup_{n\le t} U_n(s_i)$, где $U_n(s_i) = \{X_n \in E, X_{n-1} \notin E, \ldots, X_1 \notin E, X_0 = s_i\}$. Из-за дизъюнктности разных $U_n(s_i)$ из соотношения (1) получаем

$$\mathbf{P}\{X_t \in E | X_0 = \mathrm{si}\} = \Sigma_{n\le t}\mathbf{P}[U_n(s_i) | X_0 = s_i] \to 1$$

при $t \to \infty$. Так как $U_n(s_i) = \{T_i(E) = n\}$, то по σ-аддитивности получаем нужное утверждение.

Автором [10] был предложен некоторый подход для оценки средней длины траекторий Алгоритма 1 через рекуррентные соотношения.

Лемма 2. *Для любого состояния* $s_i \notin E$

$$\mathbf{E}[T_i(E)] = 1 + \Sigma^*\mathbf{E}[T_j(E)] \cdot \mathbf{P}[X_1 = s_j | X_0 = s_i], \quad (2)$$

где суммирование Σ* *идет по всем* $s_j \notin E$.

Доказательство. Из-за аддитивности среднего имеем

$$\mathbf{E}[T_i(E)] = \Sigma_{n=1}^{\infty} n \cdot \mathbf{P}[U_n(s_i) | X_0 = s_i], \quad (3)$$

где $U_n(s_i) = \{X_n \in E, X_{n-1} \notin E, \ldots, X_1 \notin E, X_0 = s_i\}$.

Тогда

$$\mathbf{E}[T_i(E)] = \Sigma_{n=1}^{\infty} n \cdot \mathbf{P}[U_n(s_i) | X_0 = s_i] = \Sigma_{n=1}^{\infty}\mathbf{P}[U_n(s_i) | X_0 = s_i] + \Sigma_{n=2}^{\infty}(n\text{-}1) \cdot \mathbf{P}[U_n(s_i) | X_0 = s_i] =$$
$$= 1 + \Sigma_{k=1}^{\infty} k \cdot \mathbf{P}[X_{k+1} \in E, X_k \notin E, \ldots, X_1 \notin E, X_0 = s_i | X_0 = s_i] =$$
$$= 1 + \Sigma_{k=1}^{\infty} k \cdot \Sigma\mathbf{P}[X_{k+1} \in E, X_k \notin E, \ldots, X_1 = s_j, X_0 = s_i | X_0 = s_i] =$$
$$= 1 + \Sigma_{k=1}^{\infty} k \cdot \Sigma^*\mathbf{P}[X_{k+1} \in E, X_k \notin E, \ldots, X_1 = s_j | X_1 = s_j] \cdot \mathbf{P}[X_1 = s_j | X_0 = s_i] = 1 +$$
$$+\Sigma_{n=1}^{\infty} n \cdot \Sigma^*\mathbf{P}[X_n \in E, X_{n\text{-}1} \notin E, \ldots, X_0 = s_j | X_0 = s_j] \cdot \mathbf{P}[X_1 = s_j | X_0 = s_i] = 1 + \Sigma^*\mathbf{E}[T_j(E)] \cdot \mathbf{P}[X_1 = s_j | X_0 = s_i].$$

Здесь мы последовательно используем тождество (3), условие марковости момента $T_i(E)$ (Теорема 1), формулу полной вероятности и однородность цепи Маркова и независимость суммы абсолютно сходящегося ряда от порядка суммирования. Этим способом совсем легко получить оценку порядка $O(n \cdot \ln n)$ на среднюю длину траектории Алгоритма 1 для n-мерной булевой алгебры. Еще более поразительный результат из [10] касается средней длины траектории Алгоритма 1 для линейных порядков. Тут верхняя граница 4 на среднюю длину не зависит от числа элементов линейного порядка!

Продемонстрируем работу этого подхода на еще одном примере.

Пример 1. Для кандидатов из обучающей выборки из Табл. 1 введем обозначения:
$\top = \langle\ \emptyset, \{f_1, f_2, f_3, f_4\}\rangle$, $o_1 = \langle\ \{o_1\}, \{f_1, f_3\}\rangle$, $o_2 = \langle\ \{o_2\}, \{f_1, f_4\}\rangle$, $o_3 = \langle\ \{o_3\}, \{f_2, f_3\}\rangle$, $o_4 = \langle\ \{o_4\}, \{f_2, f_4\}\rangle$, $f_1 = \langle\ \{o_1, o_2\}, \{f_1\}\rangle$, $f_2 = \langle\ \{o_3, o_4\}, \{f_2\}\rangle$, $f_3 = \langle\ \{o_1, o_3\}, \{f_3\}\rangle$, $f_4 = \langle\ \{o_2, o_4\}, \{f_4\}\rangle$ и $\bot = \langle\ \{o_1, o_2, o_3, o_4\}, \emptyset\ \rangle$

В этих обозначениях решетка кандидатов показана на Рис. 1. Замечательным свойством этой решетки является возможность задания расстояния между упорядоченными парами кандидатов. Расстояние 3 имеет состояние $s_0 = (\bot \le \top)$. Состояния $s_1 = (\bot \le o_1)$, …, $s_4 = (\bot \le o_4)$ и $s_5 = (f_1 \le \top)$, …, $s_8 = (f_4 \le \top)$ имеют расстояние 2. Состояния с расстоянием 1 разбиваются на две группы (средние и крайние). Крайние состояния с расстоянием 1 таковы: $s_9 = (\bot \le f_1)$, …, $s_{12} = (\bot \le f_4)$ и $s_{13} = (o_1 \le \top)$, …, $s_{16} = (o_4 \le \top)$. Средние состояния с расстоянием 1 соответствуют ребрам: $s_{17} = (o_1 \le f_1)$, …, $s_{24} = (o_4 \le f_4)$. Остальные состояния (с расстоянием 0) — эргодические.

Табл. 1. Обучающая выборка для Примера 1

$O \times F$	f_1	f_2	f_3	f_4
o_1	1	0	1	0
o_2	1	0	0	1
o_3	0	1	1	0
o_4	0	1	0	1

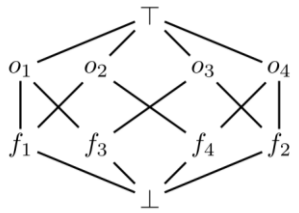

Рис. 1. Решетка кандидатов для Примера 1

Обозначим через T_3 длину траектории, стартуя из s_0. Стартовые состояния с расстоянием 2 задают длину траектории T_2. Крайние состояния расстояния 1 запускают траектории длины T_1^E, а средние – траектории длины T_1^M.

С помощью Леммы 2 выводим $\mathbf{E}T_3 = 1 + \mathbf{E}T_2$ и систему уравнений:

$$\mathbf{E}T_2 = 1 + 3/8 \cdot \mathbf{E}T_2 + 2/8 \cdot \mathbf{E}T_1^E + 2/8 \cdot \mathbf{E}T_1^M.$$
$$\mathbf{E}T_1^E = 1 + 1/8 \cdot \mathbf{E}T_2 + 2/8 \cdot \mathbf{E}T_1^E + 2/8 \cdot \mathbf{E}T_1^M.$$
$$\mathbf{E}T_1^M = 1 + 2/8 \cdot \mathbf{E}T_1^E + 2/8 \cdot \mathbf{E}T_1^M.$$

Решая эту систему, получаем, что $\mathbf{E}T_2 = 288/72 = 4$ и $\mathbf{E}T_3 = 1 + 4 = 5$.

Индуктивное обобщение обучающих примеров осуществляется следующей процедурой.

Алгоритм 2**:** Процедура индуктивного обобщения.

Data: множество обучающих (+)- и (-)-примеров; число N порождаемых гипотез
Result: случайное подмножество S гипотез объема N
O := (+)-примеры, F := признаки; $I \subseteq O \times F$ - обучающая выборка;
C := (-)-примеры, $S := \emptyset$; $i := 0$;
while ($i < N$) **do**
    Породить кандидата $\langle\ A, B\rangle$ с помощью алгоритма 1;
    hasObstacle := **false**;
    for ($c \in C$) **do**
        if ($B \subseteq \{c\}'$) **then**
            hasObstacle := **true**; **break**;
        end
    end
    if (*hasObstacle* = **false) then**
        $S := S \cup \{\langle\ A, B\rangle\ \}$; $i := i+1$;
    end
end

Проверка условия $B \subseteq \{c\}'$ в Алгоритме 2 означает, что фрагмент B кандидата $\langle\ A, B\rangle$ вкладывается в фрагмент (множество признаков) контрпримера c, т. е. (-)-примера.

Любое такое вложение означает, что кандидат нарушает условие «запрета на контр-пример». Если кандидат преодолевает все такие проверки, то он становится гипотезой (о причине наличия целевого свойства).

Процедура предсказания по аналогии (Алгоритм 3) пытается найти вложение фрагмента $B \subseteq \{o\}'$ хотя бы одной из порожденных гипотез $\langle A, B\rangle$ в фрагмент каждого тестового примера o. Если такое вложение имеется, то для этого примера предсказывается наличие целевого свойства по аналогии с родителями $A \subseteq O$ гипотезы $\langle A, B\rangle$, чей фрагмент B вложился. Иначе предсказывается отсутствие целевого свойства у этого тестового примера o.

Алгоритм 3: Процедура предсказания по аналогии.

Data: множество S гипотез, файл (?)-примеров
Result: предсказанные свойства (?)-примеров
X := (?)-примеры;
for ($o \in X$) **do**
    PredictPositively(o) := **false**;
    for ($\langle A, B\rangle \in S$) **do**
        if ($B \subseteq \{o\}'$) **then**
            PredictPositively(o) := **true**;
        end
    end
end

Для выбора числа N запусков спаривающей цепи Маркова (Алгоритма 1) в Алгоритме 2 полезно применение следующей теоремы (мы используем объекты, представленные для предсказания).

Определение 5. Тестовый пример o назовем *ε-важным*, если суммарная вероятность появления таких гипотез $\langle A, B\rangle$, что $B \subseteq \{o\}'$, будет больше ε. В работе [7] автор получил улучшенную нижнюю оценку.

Теорема 2. *Для n-признаков и любых* $\varepsilon > 0$ *и* $1 > \delta > 0$ *достаточно породить* $N \geqslant (n \cdot \ln 2 - \ln \delta)/\varepsilon$ *гипотез, чтобы с вероятностью, большей* $1 - \delta$*, все* ε*-важные тестовые примеры были предсказаны положительно.*

2. Основной результат

Перед формулировкой и доказательством основной теоремы докажем вспомогательные факты.

Определение 6. Зададим операции *CbODown* и *CbOUp* на состояниях (упорядоченных парах кандидатов) $\langle A, B\rangle \leq \langle C, D\rangle$:

$$CbODown((\langle A, B\rangle \leq \langle C, D\rangle), o) = (CbODown(\langle A, B\rangle, o) \leq CbODown(\langle C, D\rangle, o))$$

и

$$CbOUp((\langle A, B\rangle \leq \langle C, D\rangle), f) = (CbOUp(\langle A, B\rangle, f) \leq CbOUp(\langle C, D\rangle, f)).$$

Определение 7. Введем частичный *порядок между состояниями* $s_i = (\langle A_i, B_i\rangle \leq \langle C_i, D_i\rangle)$ и $s_j = (\langle A_j, B_j\rangle \leq \langle C_j, D_j\rangle)$ спаривающей цепи:

$$s_j \leqslant s_i \Leftrightarrow \langle A_i, B_i\rangle \leq \langle A_j, B_j\rangle \leq \langle C_j, D_j\rangle \leq \langle C_i, D_i\rangle.$$

Из Леммы 1 легко доказывается лемма.

Лемма 3. *Для любой упорядоченной пары состояний* $s_j \leqslant s_i$ *и любых* $o \in O$ *выполняется* $CbODown(s_j, o) \leqslant CbODown(s_i, o)$*, а для любых* $f \in F$ *верно* $CbOUp(s_j, f) \leqslant CbOUp(s_i, f)$.

Обозначим через $k = |O|$ число обучающих примеров, а через $n = |F|$ — число признаков.

Лемма 4. *Для любой упорядоченной пары невозвратных состояний* $s_j \leqslant s_i$ *спаривающей цепи Маркова выполняется* $\mathbf{E}T_j(E) \leq \mathbf{E}T_i(E)$.

Доказательство. Зададим спаренное блуждание упорядоченной пары состояний $X_t \leqslant Y_t$ формулой:

$$\mathbf{P}[X_1 = s_j', Y_1 = s_i' | X_0 = s_j, Y_0 = s_i] = m/(n + k),$$

где $m = |\{o \in O: s_j' = CbODown(s_j, o), s_i' = CbODown(s_i, o)\}| + |\{f \in F: s_j' = CbOUp(s_j, f), s_i' = CbOUp(s_i, f)\}|$

и $\mathbf{P}[X_1 = s_j', Y_1 = s_i' | X_0 = s_j, Y_0 = s_i] = 0$, если $\neg\exists\, o \in O\ [s_j' = CbODown(s_j, o), s_i' = CbODown(s_i', o)]$ & $\neg\exists\, f \in F\ [s_j' = CbOUp(s_j, f), s_i' = CbOUp(s_i, f)]$.

Из Леммы 1 следует, что $\mathbf{P}[X_1 \leqslant Y_1 | X_0 \leqslant Y_0] = 1$.

Так как из $\langle A_i, B_i\rangle = \langle C_i, D_i\rangle$ для $\langle A_i, B_i\rangle \le \langle A_j, B_j\rangle \le \langle C_j, D_j\rangle \le \langle C_i, D_i\rangle$ следует $\langle A_i, B_i\rangle = \langle A_j, B_j\rangle = \langle C_j, D_j\rangle = \langle C_i, D_i\rangle$, то с помощью Определений 4 и 7 получаем:

$$\mathbf{P}[X_t = Y_t \in E | X_0 = s_j \leqslant Y_0 = s_i] \ge \mathbf{P}[Y_t \in E | X_0 = s_j \leqslant Y_0 = s_i]. \quad (4)$$

Напомним, что для целочисленной случайной величины Z выполнено равенство $\mathbf{E}Z = \Sigma_{t=0}^{\infty} \mathbf{P}[Z > t]$. Но $X_t \notin E \Leftrightarrow T_i(E) > t$ и $Y_t \notin E \Leftrightarrow T_j(E) > t$. Из неравенств (4) следуют неравенства

$$\mathbf{P}[T_j(E) > t | X_0 = s_j, Y_0 = s_i] \le \mathbf{P}[T_i(E) > t | X_0 = s_j, Y_0 = s_i],$$

сумма которых по t завершает доказательство леммы.

Теперь рассмотрим случай, где возможно получение полиномиальной верхней границы на среднюю длину траектории Алгоритма 1. Мы будем предполагать в дальнейшем, что $O' = \emptyset$. Этого легко добиться, устраняя все признаки, общие для всех обучающих примеров.

Обогатим обучающую выборку, добавляя к каждому бинарному признаку $f_j \in F$ его отрицание $\sim f_j$. Эта конструкция часто имеет полезный смысл: мы хотим, чтобы отсутствие признака могло бы быть частью причины проявления целевого свойства.

Обогащенное множество признаков будем обозначать через F^+, а его мощность – через $2n = |F^+|$. Обычно $2n \ll k = |O|$, что мы будем предполагать в дальнейшем. Обогатим выборку $I \subseteq O \times F^+$ по правилу $o\ I\ (\sim f_j) \Leftrightarrow \neg(o\ I\ f_j)$.

Разделим все невозвратные состояния на две группы:

$V = \{s = (\langle A, B\rangle < \langle C, D\rangle): \exists f \in F^+ [f \in B]\}$ и $W = \{s = (\langle A, B\rangle < \langle C, D\rangle): \forall f \in F^+ [f \notin B]\}$.

Ясно, что состояние $s_0 = (\bot < \top) \in W$. По Лемме 4 для любого $s_j \in W$ выполнено $\mathbf{E}T_j(E) \le \mathbf{E}T_0(E)$.

По определению множества V и Лемме 4 для любого $s_j \in V$ выполнено $\mathbf{E}T_j(E) \le \mathbf{E}T_i(E)$, где $s_i = (\langle \{f\}', \{f\}''\rangle < \top) \in V$ для любого $f \in B$ при $s_j = \langle A, B\rangle$.

Введем целочисленную случайную величину Z, принимающую значение m на множестве $\{X_m = (\bot = \bot), X_{m-1} \notin V, \ldots, X_1 \notin V, X_0 = s_0\}$, которая определяет минимальное число шагов Алгоритма 1 по состояниям из $X_t \in W$ до тех пор, пока не получим $X_m = (\bot = \bot)$.

Лемма 5. *Для выборки $I \subseteq O \times F^+$ с $k = |O|$ и $|F^+| = k$ имеем* $\mathbf{E}Z \le k \cdot k! \cdot (k + 2n)/(2n)^{k+1}$.

Доказательство. События $\{Z = m\}$ возникают если будут добавлены все объекты и не будет добавлен ни один признак. Худший случай будет, если при добавлении объекта никакой другой объект не добавится в замыкание (так будет, например, в случае булевой алгебры). Если мы в нашей схеме не будем вычислять замыкание после присоединения каждого объекта, то длина цепочек объектов с повторениями, не содержащих никакого признака, которые породят все O, только увеличится.

Если $m \ge k$, то нам нужно выбрать $m!/k! \cdot (m-k)!$ способами $(m - k)$ мест, куда с вероятностью $(k/(k+2n))^{m-k}$ выберутся любые $o \in O$. Оставшиеся k мест нужно заполнить различными объектами $o \in O$. Вероятность этого $\le k!/(k+2n)^k$.

Итак, нам нужно просуммировать $\mathbf{E}Z \le k!/(k+2n)^k \cdot \Sigma_{l=0}^{\infty}(k+l) \cdot ((k+l)!/k! \cdot (l)!) \cdot (k/(k+2n))^l$. Последнюю сумму разложим на 2 слагаемых:

$$k \cdot \Sigma_{l=0}^{\infty}((k+l)!/k! \cdot (l)!) \cdot (k/(k+2n))^l + (k/(k+2n)) \cdot \Sigma_{l=0}^{\infty} l \cdot ((k+l)!/k! \cdot (l)!) \cdot (k/(k+2n))^{l-1}.$$

Воспользовавшись биномом Ньютона $(1 - x)^{-k} = \Sigma_{l=0}^{\infty}((k+l)!/k! \cdot (l)!) \cdot x^l$ и его производной $k \cdot (1 - x)^{-k-1} = \Sigma_{l=0}^{\infty} l \cdot ((k+l)!/k! \cdot (l)!) \cdot x^{l-1}$ при $x = k/(k+2n)$, получаем:

$$\mathbf{E}Z \le k!/(k+2n)^k \cdot \Sigma_{l=0}^{\infty}(k+l) \cdot ((k+l)!/k! \cdot (l)!) \cdot (k/(k+2n))^l = k \cdot k! \cdot /(2n)^k \cdot (1\text{-}k/(k+2n))^{-k} +$$
$$+ (k/(k+2n)) \cdot k! \cdot /(2n)^k \cdot (1\text{-}k/(k+2n))^{-k-1} = k \cdot k! \cdot (k + 2n)/(2n)^{k+1}.$$

Для Примера 1 $k = 4 = 2n$ и Лемма 5 дает верхнюю границу 3/4. Однако, в большинстве случаев число k объектов в обучающей выборке больше числа n признаков. Например, такая ситуация встречается в обобщении примера 1, когда $f_{2k} = \sim f_{2k-1}$ (число признаков в обогащенной матрице равно $2n$), а число примеров $k = 2^n$ для всевозможных комбинаций признаков $\{f_1, \ldots, f_{2k-1}, \ldots, f_{2n-1}\}$. При $2n \ll k$ значительно меньше оценка.

Лемма 6. *Для выборки* $I \subseteq O \times F^+$ *с* $2n = |F^+| \le k = |O|$ *имеем:*

$$\mathbf{E}Z = \Sigma_{t=1}^{\infty} \mathbf{P}[Z \ge t] \le (k+2n)\cdot(\ln 2n + (1-e^{-1})^{-1}).$$

Доказательство. Разобъем слагаемые в сумме на непересекающиеся подмножества $I_0 \cup_{r=1}^{\infty} I_r$, где $I_0 = \{1 \le l < (k+2n)\cdot \ln(2n)\}$ и $I_r = \{(k+2n)\cdot(\ln(2n) + r - 1) \le l < (k+2n)\cdot(\ln(2n)+r)\}$.

Ясно, что $\Sigma_{l=1}^{s} \mathbf{P}[Z \ge l] \le (k+2n)\cdot\ln(2n)$ для $s = (k+2n)\cdot\ln(2n)$.

Для того, чтобы произошло событие $Z \ge l$ необходимо, чтобы нашелся хотя бы один признак (из $2n$), чтобы не был выбран ни один пример в серии длины l, в котором этого признака нет. Поэтому по неравенству Буля $\mathbf{P}[Z > l] \le 2n \cdot (1 - 1/(k+2n))^l$.

Для I_r имеем $\Sigma\, \mathbf{P}[Z > l] \le 2n \cdot \Sigma\,(1 - 1/(k+2n))^l \le 2n \cdot (k+2n) \cdot (1 - 1/(k+2n))^{(k+2n)\cdot(\ln(2n)+r-1)} \le (k+2n)\cdot e^{\ln 2n} \cdot e^{-(\ln(2n)+r-1)} = (k+2n)\cdot e^{-r+1}$.

Суммируя по r, получим на $\cup_{r=1}^{\infty} I_r$ оценку $\Sigma\, \mathbf{P}[Z > l] \le (k+2n)\cdot(1-e^{-1})^{-1}$.

Обозначим наименьшую границу из Лемм 5 и 6 через *tail*.

Рассмотрим непересекающиеся события $H_l(s) = \{X_l = s \in V,\ X_{l-1} \notin V,\ \ldots,\ X_1 \notin V,\ X_0 = s_0\}$. Обозначим событие $\{X_{t+l} \in E,\ X_{t+l-1} \notin E,\ \ldots,\ X_{l+1} \notin E\} \cap H_l(s)$ через $G_{t,l}(s)$, а $\cup_{s \in V} g_{t,l}(s)$ — через $U_{t,l}$. Очевидно, имеем разложение события на дизъюнктные части:

$$\{X_{t+l} \in E,\ X_{t+l-1} \notin E,\ \ldots,\ X_1 \notin E,\ X_0 = s_0\} =$$
$$= \cup\,\{X_{t+l} \in E,\ X_{t+l-1} \notin E,\ \ldots,\ X_{l+1} \notin E\} \cap H_l(s_j) \cup \{X_{t+l} = (\perp = \perp),\ X_{t+l-1} \notin V,\ \ldots,\ X_1 \notin E,\ X_0 = s_0\} =$$
$$= U_{t,l} \cup \{X_{t+l} = (\perp = \perp),\ X_{t+l-1} \notin V,\ \ldots,\ X_1 \notin E,\ X_0 = s_0\}.$$

Нас интересует $\mathbf{E}T_0(E) = \Sigma_{l=0}^{\infty} m \cdot \mathbf{P}[X_m \in E,\ X_{m-1} \notin E,\ \ldots,\ X_1 \notin E,\ X_0 = s_0]$. Ясно, что $\mathbf{E}T_0(E) = \mathbf{E}T_0'(E) + \mathbf{E}Z$, где $\mathbf{E}T_0'(E)$ — ограничение $\mathbf{E}T_0(E)$ на $\cup_{t=1}^{\infty} \cup_{l=1}^{\infty} U_{t,l}$.

Теорема 3. *Для обогащенной выборки* $I \subseteq O \times F^+$ с $2n = |F^+| \le k = |O|$ *имеем верхнюю границу на среднюю длину траектории Алгоритма 1:*

$$\mathbf{E}T_0(E) \le (k+2n)(k^2 + k(2n+1)/2n(k^2 + k + 2n) + (k+1)(k+2n)\cdot tail/(k^2 + k + 2n). \quad (5)$$

Доказательство. Введем обозначения $R = \Sigma_{l=1}^{n}\ (1/(k+2n)) \cdot [T_{i(l)}(E) + T_{j(l)}(E)]$, где $s_{i(l)} = (\langle\ \{f_l\}',\ (\{f_l\}')'\rangle\ <\top)$, аналогично для $s_{j(l)} = (\langle\ \{\sim f_l\}',\ (\{\sim f_l\}')'\rangle\ <\top)$. Тогда по марковскому свойству и однородности:

$$\mathbf{E}T_0'(E) = \Sigma_{t=1}^{\infty}\Sigma_{l=1}^{\infty}(t+l)\cdot\mathbf{P}[U_{t,l}] = \Sigma_{t=1}^{\infty}\Sigma_{s \in V}\, t \cdot \mathbf{P}[\{X_t \in E,\ X_{t-1} \notin E,\ \ldots,\ X_1 \notin E |$$
$$X_0 = s\}] \cdot \Sigma_{l=1}^{\infty}\mathbf{P}[H_l(s)] +$$
$$\Sigma_{l=1}^{\infty}\Sigma_{s \in V}\, l \cdot \mathbf{P}[H_l(s)] \cdot \Sigma_{t=1}^{\infty}\mathbf{P}[\{X_t \in E,\ X_{t-1} \notin E,\ \ldots,\ X_1 \notin E |\ X_0 = s\}] \le$$
$$\le \Sigma_{s \in V}\mathbf{E}T_s(E) \cdot \mathbf{P}[X_1 = s |\ X_0 = s_0] + \Sigma_{s \in V}\Sigma_{l=1}^{\infty}\, l \cdot \mathbf{P}[H_l(s)] =$$
$$= \Sigma_{s \in V}\mathbf{E}T_s(E) \cdot \mathbf{P}[X_1 = s |\ X_0 = s_0] + (k+2n)/2n,$$

где последнее слагаемое – математическое ожидание геометрически распределенной случайной величины ожидания выбора первого признака.

По формуле полной вероятности и Лемме 4 имеем:

$$\mathbf{E}T_{i(l)}(E) \le 1 + \Sigma_{m=1}^{n}\,(1/(k+2n)) \cdot [\mathbf{E}T_{i(m)}(E) + \mathbf{E}T_{j(m)}(E)] - (1/(k+2n)) \cdot \mathbf{E}T_{j(l)}(E) + k/(k+2n) \cdot \mathbf{E}T_0(E).$$

Отсюда следует: $\mathbf{E}R \le 2n/(k+2n) \cdot [1 + \mathbf{E}R + k/(k+2n) \cdot \mathbf{E}T_0(E)] - (1/(k+2n)) \cdot \mathbf{E}R$. Поэтому

$$(k+1) \cdot \mathbf{E}R/(k+2n) \le 2n/(k+2n) + 2kn/(k+2n)^2 \cdot \mathbf{E}T_0(E).$$

Подставив $\mathbf{E}R \le 2n/(k+1) + 2kn \cdot \mathbf{E}T_0(E)/(k+1)(k+2n)$ в неравенство $\mathbf{E}T_0(E) \le \mathbf{E}R + (k+2n)/2n + \mathbf{E}Z$, получим $(k^2 + k + 2n)\cdot\mathbf{E}T_0(E)/(k+1)(k+2n) \le 2n/(k+1) + (k+2n)/2n + tail$, что и приводит к нужной оценке.

Заключение

Статья содержит описание значительного продвижения в решении открытой проблемы ВКФ-метода [9] о нахождении полиномиальной верхней границы на среднее время склеивания для спаривающей цепи Маркова–средней длины траектории вероятностного алгоритма порождения кандидатов в гипотезы о причинах целевого свойства. Соединяя описанный результат с ранее полученной полиномиальной нижней оценкой на достаточное число гипотез, получаем полностью полиномиальную схему извлечения знаний с помощью бинарной операции сходства, реализованную в ВКФ-методе.

Автор благодарен своим коллегам по ВЦ им. А. А. Дородницына Федерального исследовательского центра «Информатика и управление» РАН, а также проф. С. О. Кузнецову за полезные обсуждения и поддержку. Особо хочется поблагодарить аспирантку Центра Л. А. Якимову за совместную работу, побуждающую автора теоретически обосновывать интуицию, сформированную по результатам обсуждения ее экспериментов над реальными данными, однако за все неточности и ошибки, как обычно, несет ответственность исключительно автор.

Литература

1. ДСМ-метод автоматического порождения гипотез: Логические и эпистемологические основания //Ред.: Финн В.К., Аншаков О.М.). М.: URSS. 2009. 432 с.
2. Милль Дж.Ст. Система логики силлогистической и индуктивной: Изложение принципов доказательства в связи с методами научного исследования. Пер. с англ. Изд. 5. М.: URSS. 2011. 832 с.
3. Гусакова С.М., Финн В.К. Сходства и правдоподобный вывод // Известия АН СССР. Сер. «Техническая кибернетика». 1987. № 5. С. 42–63.
4. Ganter, Bernhard and Wille, Rudolf. Formal Concept Analysis: Mathematical Foundations. Berlin: Springer–Verlag. 1999. 284 p.
5. Виноградов Д.В. Вероятностное порождения гипотез в ДСМ-методе с помощью простейших цепей Маркова // Научная и техническая информация. Сер. 2. 2012. № 9. С. 20–27.
6. Кузнецов С.О. Быстрый алгоритм построения всех пересечений объектов из нижней полурешетки // Научная и техническая информация. Сер. 2. 1993. № 1. С. 17–20.
7. Виноградов Д.В. Алгебраическое машинное обучение: упор на эффективность // Автоматика и телемеханика. 2022. № 6. С. 5–23.
8. Кемени Дж., Снелл Дж. Конечные цепи Маркова. Пер. с англ. М.: Наука. 1970. 272 с.
9. Виноградов Д.В. ВКФ-метод интеллектуального анализа данных: обзор результатов и открытых проблем // Искусственный интеллект и принятие решений. 2017. № 2. С. 9–16.
10. Виноградов Д.В. Цепи Маркова, формула полной вероятности и рекуррентные соотношения // Научная и техническая информация. Сер. 2. 2023. № 2. С. 35–39.

Виноградов Дмитрий Вячеславович. Доктор физико-математических наук. Ведущий научный сотрудник. Федеральный исследовательский центр «Информатика и управление» Российской академии наук. Области исследований: вероятностные алгоритмы, дискретная математика, математическая логика. E-mail: KRRGuest@yandex.ru

On Computational Efficiency of Knowledge Extraction by Probabilistic Algorithms

D. V. Vinogradov

Federal Research Center «Computer Science and Control» RAS, Moscow, Russia

Abstract. The paper demonstrates computational efficiency of probabilistic approach to knowledge extraction through binary similarity operation. In addition to previously proved by the author the result on sufficiency of a polynomial number of hypotheses on causes of investigated target property, the paper contains a polynomial upper bound on mean working time of the algorithm to generate a single candidate for hypothesis. The proven result concerns a family of algorithms based on coupled Markov chains. To obtain a good estimate for the length of the trajectory (before entering the ergodic state) of such a chain, we needed to enrich the training sample by adding negative columns for existing binary features.

Keywords: similarity, candidate, coupled Markov chain, average length of trajectory.

DOI 10.14357/20718594230403 **EDN** NELPQW

References

1. Finn V.K., Anshakov O.M. DSM-metod avtomaticheskogo porozhdeniya gipotez: Logicheskie i epistemologicheskie osnovaniya [JSM Method for Automatic Hypotheses Generation: Logical and Epistemological Foundations]. Moscow: Editorial URSS, 2009.
2. Mill J.S. A System of Logic. Honolulu: University Press of the Pacific, 2002.
3. Gusakova S.M., Finn V.K. Shodstva i pravdopodobnyj vyvod [Similarities and Plausible Inference]. Izvestia AN SSSR, Ser. «Technical cybernetics». 1987. No 5. P. 42–63.
4. Ganter B., Wille R. Formal Concept Analysis: Mathematical Foundations. Berlin: Springer, 1999.

5. Vinogradov D.V. Random generation of hypotheses in the JSM method using simple Markov chains. Automatic Documentation and Mathematical Linguistics. 2012. No 46(5). P. 221–228.
6. Kuznetsov S.O. A Fast Algorithm for Computing All Intersections of Objects in a Finite Semi-Lattice, Automatic Documentation and Mathematical Linguistics. 1993. No 27(5). P. 11-21.
7. Vinogradov D.V. Algebraic Machine Learning: Emphasis on Efficiency. Automation and Remote Control. 2022. No 83(6). P. 831–846.
8. Kemeny J.G., Snell J.L. Finite Markov chains. New York: Springer, 1976.
9. Vinogradov D.V. The VKF Method for Data Mining: a Survey of the State of the Art and Open Problems. Scientific and Technical Information Processing. 2018. No 45(6). P. 411–416.
10. Vinogradov D.V. Markov Chains, Law of Total Probability, and Recurrence Relations. Automatic Documentation and Mathematical Linguistics. 2023. No 57(1). P. 68–72.

Vinogradov Dmitry V. Doctor of physical and mathematical sciences. Leading Researcher, Federal Research Center “Computer Science and Control”, Russian Academy of Sciences. Research areas: probabilistic algorithms, discrete mathematics, mathematical logic. Email: KRRGuest@yandex.ru