УДК 519.217

Д.В. Виноградов

Субмультипликативность и остановка спаривающей цепи Маркова для ВКФ-метода*

Исследуется вариант спаривающей цепи Маркова для ВКФ-метода, где предлагается останавливать излишне длинную траекторию, если число шагов в ней превосходит сумму длин траекторий, вычисленных заранее. Для этого варианта алгоритма доказывается лемма о субмультипликативности. Доказывается теорема о вероятности длины траектории превзойти заданный порог с помощью рассуждения в духе леммы Фекете. Наконец, доказывается, что вероятности результатов обычной и останавливаемой спаривающих цепей Маркова отличаются в метрике тотальной вариации на экспоненциально малую величину от числа учитываемых предварительных траекторий.

Ключевые слова: *спаривающая цепь Маркова, остановка траектории, субмультипликативность, лемма Фекете, метрика тотальной вариации*

ВВЕДЕНИЕ

Вероятностно-комбинаторный* подход к машинному обучению, основанному на бинарной операции сходства, был нами предложен в работе [1]. Многие технические факты о решетке сходств взяты из анализа формальных понятий [2], поэтому этот подход называется вероятностно-комбинаторным формальным методом, сокращенно ВКФ-методом.

Ключевой процедурой ВКФ-метода является вероятностный алгоритм нахождения сходств с помощью спаривающей цепи Маркова [3]. Нами показано, что это – действительно цепь Маркова, которая останавливается с вероятностью единица. Конечность траекторий доказывается как следствие классической теоремы о невозвратных состояниях счетной цепи Маркова [4]. В одном важном частном случае (Булева алгебра всех подмножеств) нами [5] была получена точная формула для средней длины траекторий (порядка) и доказана теорема о сильной концентрации длин траекторий около своего среднего. Однако хорошей (полиномиальной) оценки на среднюю длину траекторий в общем случае получить пока не удалось.

Как показывает случай Булевой алгебры, хотя подавляющее число траекторий имеет полиномиальную длину, не исключена возможность, что малое число траекторий может иметь экспоненциальную длину. Тогда можно предложить использовать r предварительных прогонов цепи Маркова для получения верхней границы на момент остановки (как суммы длин этих предварительно вычисленных траекторий). Если текущая траектория не завершается до этой границы, то текущие вычисления прекращаются, а цепь Маркова запускается заново. При такой модификации изменяются вероятности появления результатов (итоговых сходств).

В настоящей работе будет получена хорошая оценка, показывающая, что при выборе достаточно большого числа предварительных прогонов, изменение вероятностей может быть сделано сколь угодно малым, а в качестве технического средства будет предварительно установлен закон субмультипликативности и доказан некоторый вариант леммы Фекете (см., например, [6]).

СПАРИВАЮЩАЯ ЦЕПЬ МАРКОВА И СУБМУЛЬТИПЛИКАТИВНОСТЬ

Фундаментальная теорема анализа формальных понятий [2] утверждает, что любую (нижнюю полу) решетку сходств можно изоморфно заменить (полу)решеткой битовых строк с операцией побитового умножения в качестве сходства. Алгоритм кодирования объектов битовыми строками для формирования минимального (формального) контекста [2], порождающий изоморфную решетку на битовых строках, был предложен в статье [7].

Контекст – это бинарное отношение между элементами множества O, которые мы называем *объектами*, и элементами множества F, которые мы называем *признаками*. Если в строке, соответствующей объекту $o \in O$, и столбце, соответствующим фрагменту $f \in F$, стоит единица, то мы говорим, что *объект* ***o*** *обладает признаком* ***f***, и обозначаем это через ***olf***. В противном случае, говорим, что *объект* ***o*** *не имеет признака* ***f***.

Для подмножества объектов его *сходством* называется подмножество $A' = \{f \in F : \forall o \in A[olf]\} \subseteq F$. Договорились, что $\varnothing' = F$.

* Работа выполнена при частичной поддержке гранта РФФИ № 18-29-03063мк

На самом деле, это определение совпадает с последовательным вычислением побитового умножения строк, соответствующих отобранным во множество A объектов.

Для подмножества $B \subseteq F$ признаков его *сходством* называется подмножество

$$B' = \{o \in O : \forall f \in B[olf]\} \subseteq O .$$

Условились считать, что $\varnothing' = O$.

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

Определение 2. Операция ***замыкай-по-одному-вниз*** на кандидате $\langle A,B \rangle$ и объекте $o \in O$ порождает пару $CbODown(\langle A,B \rangle, o) = \left\langle (A \cup \{o\})'', (A \cup \{o\})' \right\rangle$.

Операция ***замыкай-по-одному-вверх*** на кандидате $\langle A,B \rangle$ и признаке $f \in F$ порождает пару $CbOUp(\langle A,B \rangle, f) = \left\langle (B \cup \{f\})', (B \cup \{f\})'' \right\rangle$.

Основным объектом изучения в настоящей работе будет следующий алгоритм:

Data: множество обучающих (+)-примеров; внешние функции *CbOUp*(,) и *CbODown*(,) операций «замыкай-по-одному»
Result: кандидат $\langle A,B \rangle$
O := (+)-примеры, F := признаки; $I \subseteq O \times F$ – формальный контекст для (+)-примеров;
$R := O \cup F$; $Min := \langle O, O' \rangle$; $Max := \langle F', F \rangle$;
while ($Min \neq Max$) **do**
    Выбираем случайный элемент $r \in R$;
    if ($r \in O$) **then**
        $Min := CbODown(Min, r)$;
        $Max := CbODown(Max, r)$;
    end
    else
        $Min := CbOUp(Min, r)$;
        $Max := CbOUp(Max, r)$;
    end
end

Алгоритм 1: Спаривающая цепь Маркова

Определение 3. ***Порядок*** на кандидатах зададим правилом $\langle A_1, B_1 \rangle \le \langle A_2, B_2 \rangle$, если $B_1 \subseteq B_2$. В анализе формальных понятий [2] определение порядка задается двойственным образом. Наше определение соответствует традиции отечественной школы.

Заметим, что состоянием изменяемых переменных в цикле Алгоритма 1 (= состоянием спаривающей цепи Маркова) является упорядоченная пара кандидатов $\langle A_1, B_1 \rangle \le \langle A_2, B_2 \rangle$.

Первоначально меньший кандидат совпадает с наименьшим кандидатом $Min := \langle O, O' \rangle$, а больший – с наибольшим $Max := \langle F', F \rangle$.

В цикле Алгоритма 1 к обоим кандидатам применяется одна и та же операция *CbODown* с выбранным объектом, или операция *CbOUp* с выбранным признаком.

Процесс останавливается, когда меньший кандидат совпадет (склеится) с большим. Тогда этот общий кандидат и выдается Алгоритмом 1 в качестве результата.

Для дальнейших доказательств полезна легко проверяемая

Лемма 1. *Для упорядоченной пары кандидатов* $\langle A_1, B_1 \rangle \le \langle A_2, B_2 \rangle$ *и любого объекта* $o \in O$ *имеем* $CbODown(\langle A_1, B_1 \rangle, o) \le CbODown(\langle A_2, B_2 \rangle, o)$.

Для упорядоченной пары кандидатов

$$\langle A_1, B_1 \rangle \le \langle A_2, B_2 \rangle$$

и любого признака $f \in F$ *выполнено*

$$CbOUp(\langle A_1, B_1 \rangle, f) \le CbOUp(\langle A_2, B_2 \rangle, f).$$

В работе [1] с помощью Леммы 1 была доказана

Теорема 1. *Алгоритм 1 соответствует цепи Маркова.*

Определение 4. Состояние вида $\langle A_1, B_1 \rangle < \langle A_2, B_2 \rangle$ называется ***невозвратным***.

Классической теоремой о счетных цепях Маркова (см., например, [4]) является

Теорема 2. *Вероятность того, что состояние* $\langle A_1(t), B_1(t) \rangle \le \langle A_2(t), B_2(t) \rangle$ *спаривающей цепи Маркова в момент времени t окажется невозвратным, стремится к нулю, когда* $t \to \infty$.

Отсюда следует, что Алгоритм 1 будет останавливаться с вероятностью 1.

Определение 5. Первый шаг T спаривающей цепи Маркова, когда $\langle A_1(T), B_1(T) \rangle = \langle A_2(T), B_2(T) \rangle$ называется ***моментом склеивания***.

Установим субмультипликативность $\mathbf{P}[T > t]$.

Лемма 2. *Для любых натуральных t и s выполнено* $\mathbf{P}[T > t+s] \le \mathbf{P}[T > t] \cdot \mathbf{P}[T > s]$.

Доказательство следует из формулы условной вероятности, так как если $[T > s]$, то

$$\langle A_1(s), B_1(s) \rangle < \langle A_2(s), B_2(s) \rangle$$

Поэтому, применяя ко всем четырем ВКФ-кандидатам

$$\langle A_3(\tau), B_3(\tau) \rangle \le \langle A_1(\tau+s), B_1(\tau+s) \rangle <$$
$$< \langle A_2(\tau+s), B_2(\tau+s) \rangle \le \langle A_4(\tau), B_4(\tau) \rangle$$

одинаковые операции *CbODown* и *CbOUp*, где $\langle A_3(0), B_3(0) \rangle := \langle O, O' \rangle$ и $\langle A_4(0), B_4(0) \rangle := \langle F', F \rangle$, имеем по Лемме 1, что если

$$\langle A_1(\tau+s), B_1(\tau+s) \rangle < \langle A_2(\tau+s), B_2(\tau+s) \rangle,$$

т.е. исходная пара склеивается позднее момента $t+s$, то и $\langle A_3(t), B_3(t) \rangle < \langle A_4(t), B_4(t) \rangle$, т.е. склеивание совершается позднее момента *t*. Это и означает, что $\mathbf{P}(T > t+s | T > s) \le \mathbf{P}[T > t]$.

Теперь мы применим полученную субмультипликативность для асимптотической оценки вероятности того, что длина траектории превзойдет заданный порог.

Теорема 3. *Найдется такое число* $-\infty \le \gamma < 0$,

чтобы $\lim_{t\to\infty} \frac{\ln \mathbf{P}[T>t]}{\tau} = \gamma$.

Доказательство. Положим

$$inf_{t\to\infty} \frac{\ln \mathbf{P}[T>t]}{\tau} = \gamma \le 0.$$

Случай $\gamma = 0$ невозможен, так как иначе для любого t будет $\mathbf{P}[T>t] \ge 1$, что противоречит Теореме 2 из-за известного равенства $\mathbf{E}[T] = \sum_{t=0}^{\infty} \mathbf{P}[T>t]$ для целочисленной случайной величины. Теперь рассмотрим случай $\gamma > -\infty$. Для любого t выполняется $\ln \mathbf{P}[T>t] \ge \gamma \cdot t$ и для любого $\varepsilon > 0$ найдется такое s, что $\ln \mathbf{P}[T>t] \le (\gamma+\varepsilon)\cdot s$. Деля t с остатком на s, получаем $t = q\cdot s + r$, где $0 \le r < s$. По Определению 5, $\mathbf{P}[T > q\cdot s] \ge \mathbf{P}[T > t = q\cdot s + r]$ откуда по Лемме 2 следует, что

$$\gamma + \varepsilon = \lim_{q\to\infty} \frac{(\gamma+\varepsilon)\cdot q\cdot s}{q\cdot s + r} \ge \limsup_{t\to\infty} \frac{\ln \mathbf{P}[T>t]}{t} \ge \gamma.$$

Из-за произвольности $\varepsilon > 0$ доказательство для случая $\gamma > -\infty$ закончено. Наконец, рассмотрим случай $\gamma = -\infty$. Тогда для любого $\delta < 0$ найдется такое s, что $\ln \mathbf{P}[T>s] \le \delta \cdot s$. Деля t с остатком на s, получаем $t = q\cdot s + r$, где $0 \le r < s$. По Определению 5 $\mathbf{P}[T > q\cdot s] \ge \mathbf{P}[T > t = q\cdot s + r]$, откуда по Лемме 2 следует, что

$$0 > \delta = \lim_{q\to\infty} \frac{\delta\cdot q\cdot s}{q\cdot s + r} \ge \limsup_{t\to\infty} \frac{\ln \mathbf{P}[T>t]}{t}.$$

Из-за произвольности $\delta < 0$ доказательство для случая $\gamma = -\infty$ закончено.

ОСТАНОВЛЕННАЯ ЦЕПЬ МАРКОВА

Так как спаривающая цепь Маркова может иметь траектории существенно разной длины, то возможно применение следующей техники остановки длинной траектории и запуска цепи заново:

Определение 6. Если $T_1,...,T_r$ – независимые целочисленные случайные величины, имеющие распределение времени T склеивания, то ***верхняя граница склеивания*** по r предварительным прогонам определяется как $\hat{T} = T_1 + ... + T_r$.

На практике предлагается сделать r прогонов спаривающей цепи Маркова с соответствующими временами склеивания $t_1,...,t_r$ и взять оценку $t_1 + ... + t_r$ верхней границы склеивания.

Определение 7. Для целочисленной случайной величины $\hat{T}$, независимой от целочисленной случайной величины T, ***условное распределение*** состояний относительно события $B = \left\{T \le \hat{T}\right\}$ есть распределение

$$\mu\left(\hat{T}\right)_i = \frac{\mathbf{P}\left[X_T = i, T \le \hat{T}\right]}{\mathbf{P}\left[T \le \hat{T}\right]}$$

для любого состояния i.

Определение 8. Расстояние ***тотальной вариации*** между распределениями вероятностей $\mu = (\mu_i)_{i\in U}$ и $\nu = (\nu_i)_{i\in U}$ на конечном пространстве U определяется правилом

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\cdot \sum\nolimits_{i\in U} |\mu_i - \nu_i|.$$

Это расстояние является половиной метрики l_1, следовательно, само является метрикой (в частности, симметрично).

Следующая лемма является технической.

Лемма 3. $$\left\|\mu - \mu\left(\hat{T}\right)\right\|_{TV} \le \frac{\mathbf{P}\left[T > \hat{T}\right]}{1 - \mathbf{P}\left[T > \hat{T}\right]},$$

где $\mu\left(\hat{T}\right)$ *– распределение остановленной на верхней границе* $\hat{T}$ *склеивания по* $r<1$ *испытаниям, а* μ *– распределение выдачи неостановленной цепи.*

Доказательство. По определению 8

$$\mu\left(\hat{T}\right)_i = \frac{\mathbf{P}\left[X_T = i, T \le \hat{T}\right]}{\mathbf{P}\left[T \le \hat{T}\right]}.$$

Тогда

$$\mathbf{P}\left[T \le \hat{T}\right]\cdot\left(\mu\left(\hat{T}\right)_i - \mu_i\right) = \mathbf{P}\left[X_T = i, T \le \hat{T}\right] - \mathbf{P}\left[T \le \hat{T}\right]\cdot\mu_i =$$
$$\mathbf{P}\left[T > \hat{T}\right]\cdot\mu_i - \mathbf{P}\left[X_T = i, T > \hat{T}\right] \le \mathbf{P}\left[T > \hat{T}\right]\cdot\mu_i.$$

Суммируя по множеству $R = \left\{i \in U \middle| \mu_i > \mu\left(\hat{T}\right)_i\right\}$, получим

$$\mathbf{P}\left[T \le \hat{T}\right]\cdot\left\|\mu - \mu\left(\hat{T}\right)\right\|_{TV} \le \mathbf{P}\left[T > \hat{T}\right],$$

что и приводит к утверждению леммы.

Теперь докажем основной результат.

Теорема 4. *Имеет место неравенство*

$$\left\|\mu - \mu\left(\hat{T}\right)\right\|_{TV} \le \frac{1}{2^r - 1},$$

где $\mu\left(\hat{T}\right)$ *– распределение остановленной на верхней границе* $\hat{T}$ *склеивания по* $r<1$ *испытаниям, а* μ *– распределение выдачи неостановленной цепи.*

Доказательство. Докажем сначала, что

$$\mathbf{P}\left[T > \hat{T}\right] \le 2^{-r}.$$

Из определения $T,T_1,...,T_r$ как независимых одинаково распределенных случайных величин, следует, что $\mathbf{P}\left[T > T_j\right] \le \frac{1}{2}$ для всех $1 \le j \le r$. С помощью Леммы 2 доказывается неравенство

$$\mathbf{P}\left[T > \sum_{j=1}^{t} T_j\right] \le \mathbf{P}\left[T > \sum_{j=1}^{r-1} T_j\right] \cdot \mathbf{P}[T > T_r] \le ... \le$$
$$\le \prod_{j=1}^{r} \mathbf{P}\left[T > T_j\right] \le 2^{-r}$$

По Лемме 3 имеем

$$\left\|\mu - \mu\left(\hat{T}\right)\right\|_{TV} \le \frac{\mathbf{P}\left[T > \hat{T}\right]}{1 - \mathbf{P}\left[T > \hat{T}\right]} = \frac{2^{-r}}{1 - 2^{-r}} = \frac{1}{2^{-r} - 1}.$$

Известна классическая и доказываемая прямо из Определения 8

Лемма 4. *Выполняется равенство*

$$\left\|\mu - \nu\right\|_{TV} = \mathbf{max}_{R \subseteq U} \left|\mu(R) - \nu(R)\right|.$$

В Лемме 4 подмножество *R*, на котором достигается максимум, определяется так: $R = \left\{i \in U \middle| \mu_i > \nu_i\right\}$.

Соединяя результаты Леммы 4 и Теоремы 4, получим

Следствие 1. *Для любого подмножества состояний R с* $\mu(R) = \rho$, *если взять число предварительных запусков равным* $r > \log_2\left(1 - \frac{1}{\rho}\right)$, *то имеем* $\mu\left(\hat{T}\right)(R) \ge \rho - \frac{1}{2^r - 1}$ *для цепи Маркова, остановленной по верхней границе* $\hat{T}$ *склеивания по r испытаниям.*

Доказательство.

$$\rho - \frac{1}{2^r - 1} \le \mu(R) - \left\|\mu - \mu\left(\hat{T}\right)\right\|_{TV} =$$
$$= \mu(R) - \mathbf{max}_{Q \subseteq U} \left|\mu(Q) - \mu\left(\hat{T}\right)(Q)\right| \le$$
$$\le \mu(R) - \left|\mu(R) - \mu\left(\hat{T}\right)(R)\right| \le \mu\left(\hat{T}\right)(R).$$

ЗАКЛЮЧЕНИЕ

Нами был исследован один механизм ускорения вычисления сходств в ВКФ-методе с использованием спаривающей цепи Маркова. Предлагается использовать предварительно вычисленные траектории цепи Маркова для получения верхней границы на момент остановки (как суммы длин этих траекторий). Если текущая траектория не завершается до этой границы, то текущие вычисления прекращаются, а цепь Маркова запускается заново.

В настоящей работе получена хорошая оценка, показывающая, что при выборе достаточно большого числа предварительных прогонов (неостановленной) спаривающей цепи Маркова, изменение вероятностей выдач результатов остановленной цепью Маркова относительно стандартной спаривающей цепи Маркова может быть сделано экспоненциально малым.

Предварительно был установлен закон субмультипликативности для вероятностей превышения порога длиной траектории спаривающей цепи Маркова и с его помощью доказана теорема о вероятности для длины траектории превзойти заданный порог.

* * *

Автор благодарит своих коллег в ФИЦ ИУ РАН и РГГУ за поддержку и полезные дискуссии. Особая благодарность выражается студентам отделения интеллектуальных систем РГГУ, которые выступали первыми слушателями и критиками описываемого подхода (в рамках курса «Интеллектуальный анализ данных и машинное обучение»).

СПИСОК ЛИТЕРАТУРЫ

1. Vinogradov D.V. VKF-method of hypotheses generation // Communications in Computer and Information Science. – 2014. – Vol. 436. – P. 237-248
2. Ganter B., Wille R. Formal Concept Analysis. Transl. from German. – Berlin: Springer–Verlag, 1999. – 284 p.
3. 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; Vinogradov D.V. Random Generation of Hypotheses in the JSM Method using Simple Markov Chains // Automatic Documentation and Mathematical Linguistics. – 2012. – Vol. 46, № 5. – P. 221-228.
4. Кемени Дж., Снелл Дж., Кнэпп А. Счетные цепи Маркова / пер. с англ. – М.: Наука, 1987. – 416 с.
5. Виноградов Д.В. Анализ результатов применения ВКФ-системы: успехи и открытая проблема // Научная и техническая информация. Сер. 2. – 2017. – № 5. – С. 1-4; Vinogradov D.V. Analysis of the Results of Application of the VKF System: Successes and an Open Problem // Automatic Documentation and Mathematical Linguistics. – 2017. – Vol. 51, № 3. – P. 108-111.
6. Steel J. Michael. Probability Theory and Combinatorial Optimization, CBMS-NSF regional conference series in applied mathematics. Vol. 69 – Philadelphia (PA): SIAM, 1998. – 159 p.
7. Виноградов Д.В. О представлении объектов битовыми строками для ВКФ-метода // Научная и техническая информация. Сер. 2. – 2018. – № 5. – С. 1-4; Vinogradov D.V. On Object Representation by Bit Strings for the VKF-Method // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 3. – P. 113-116.

Материал поступил в редакцию 10.08.18

Сведения об авторе

ВИНОГРАДОВ Дмитрий Вячеславович – кандидат физико-математических наук, старший научный сотрудник Федерального исследовательского центра «Информатика и управление» РАН и доцент Российского Государственного Гуманитарного Университета, Москва
e-mail: vinogradov.d.w@gmail.com