УДК 519.217.2

А.М. ГУПАЛ, Н.А. ГУПАЛ

СВОЙСТВА ПРОЦЕДУР СЕПАРАЦИИ ДЛЯ ДИСКРЕТНЫХ ОБЪЕКТОВ В МОДЕЛЯХ БАЙЕСОВСКИХ СЕТЕЙ

Ключевые слова: *сепарация, байесовская сеть, условие Маркова, дискретный объект.*

ВВЕДЕНИЕ

В теории байесовских сетей разработаны оригинальные процедуры сепарации (d-сепарации) для определения множества условных независимостей между переменными (вершинами) ациклически ориентированного графа. Сепарационный подход использован в алгоритмах вывода структуры графа из эмпирических данных. Особенность методов сепарации заключается в том, что они опираются лишь на структуру графа, а условные независимости формулируются в терминах вероятностных распределений. Установлена связь между d-сепарацией и условной независимостью: при выполнении условия Маркова каждая d-сепарация является условной независимостью. Обратное утверждение о том, что каждая условная независимость определяется d-сепарацией, выполняется при дополнительном жестком предположении достоверности (faithfulness) — когда из условия Маркова выводятся все условные независимости между переменными модели [1, 2].

В линейном случае, когда каждая переменная является функцией от родительских переменных и случайных помех и множество возможных условных независимостей образует вещественное пространство, показано, что множество распределений, для которых условие достоверности не выполняется, образует лебеговское множество меры нуль, т.е. почти все условные независимости определяются d-сепарацией [2].

В настоящей работе для дискретных объектов, построенных на ограниченных выборках, приведены примеры, подтверждающие предположение достоверности, и примеры (их приблизительно столько же), для которых оно не выполняется. Таким образом, ряд теорем о сепарационном подходе в [1–3] и других монографиях по байесовским сетям теряют смысл для дискретных объектов, в том числе и само предположение достоверности.

БАЗОВЫЕ ОПРЕДЕЛЕНИЯ

Ориентируемый граф – пара $\mathbf{G} = (\mathbf{V}, \mathrm{E})$, где $\mathbf{V}$ — конечное множество вершин (переменных), E — множество ориентированных пар элементов из $\mathbf{V}$, называемых дугами. Считаем, что каждая переменная $X \in \mathbf{V}$ принимает конечное число значений $\{x\}$. Если в графе есть дуга $X \to Y$, то вершина x называется «родителем» вершины y, $F(x)$ — множество родителей x. Когда ориентация дуги неизвестна или игнорируется, она обозначается как ребро: $x - y$. Вершины, соединенные ребром, называются смежными, а путь, на котором все ребра ориентированы в одном направлении $x \to \ldots \to y$, — орпутем. Вершина Y называется потомком вершины X, если существует орпуть из X в Y. Фрагмент вида $x \to y \leftarrow z$ называется коллайдером. Ациклический ориентированный граф (АОГ) — это орграф, в котором ориентированные циклы отсутствуют. Термины «переменная» и «вершина» употребляются как взаимозаменяемые.

Пара $(\mathbf{G}, P)$ удовлетворяет условию Маркова, если для каждой переменной $X \in \mathbf{V}$ множество $\{x\}$ условно независимо от множества всех непотомков $\mathbf{ND}(x)$ при заданном множестве родителей и обозначается $I_P(\{x\}, \mathbf{ND}(x) | F(x))$, где P — вероятностное распределение. Если $(\mathbf{G}, P)$ удовлетворяет условию Маркова, то

$$P(x_1, \ldots, x_n) = \prod_i P(x_i | F(x_i)), \tag{1}$$

т.е. вероятность P равна произведению условных вероятностей всех переменных при заданных значениях их родителей.

КРИТЕРИЙ d-СЕПАРАЦИИ

Пусть Z — множество вершин, X и Y — вершины во множестве $\mathbf{V} - Z$. Путем между вершинами X и Y называют последовательность дуг из X в Y произвольной ориентации. Путь ρ между вершинами X и Y блокируется множеством Z, если выполняется одно из следующих условий:

– путь ρ содержит цепочку $i \to m \to j$ или ветвление $i \leftarrow m \to j$ такие, что $m \in Z$;

– путь ρ содержит коллайдер $\to u \leftarrow$, причем $u \notin Z$ и нет никакой вершины $w \in Z$ такой, что существует орпуть $u \to \ldots \to w$.

Множество Z d-сепарирует вершины X и Y, если Z блокирует все пути между X и Y. Таким образом, d-сепарация блокирует поток информации между вершинами X и Y. Этот факт обозначается $I_{\mathbf{G}}(X, Y | Z)$. Естественным образом d-сепарация распространяется на множества вершин X и Y.

В [2] показано, что из d-сепарации вытекает условная независимость.

Теорема 1. Если вероятностное распределение P удовлетворяет условию Маркова для графа $\mathbf{G}$ и множество Z d-сепарирует множества вершин X и Y, то X и Y независимы при заданном множестве Z, т.е. имеет место $I_P(X, Y | Z)$.

Отметим, что d-сепарация — дискретная процедура на графе. Ее возможности ограничены структурой графа, т.е. количеством вершин и дуг. Обратное утверждение к теореме 1 выполняется при достаточно сильном предположении о достоверности того, что из условия Маркова выводятся все условные независимости между переменными. Данное предположение проверить невозможно, поскольку структура графа заранее неизвестна, а существует, как правило, ограниченное множество статистических данных.

Теорема 2. Вероятностное распределение $P(\mathbf{V})$ удовлетворяет условию достоверности, если и только если X и Y независимы при заданном множестве Z, если и только если множество Z d-сепарирует X и Y.

Теорема 2 выражает эквивалентность процедуры d-сепарации и множества условных независимостей при предположении о достоверности, которое может не выполняться на некоторых вероятностных распределениях.

В монографии [3] приведены ошибочные утверждения относительно свойств d-сепарации, в частности лемма 2.3 и теорема 2.1, в которой утверждается эквивалентность процедуры d-сепарации и условной независимости без предположения о достоверности. Рассмотрим эти ошибки.

Лемма 2.3 из [3]. Пусть вероятностное распределение P удовлетворяет условию Маркова для графа $\mathbf{G}$ и множества $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$ взаимно не пересекаются. Тогда $I_P(\mathbf{A}, \mathbf{B} | \mathbf{C}) \to I_{\mathbf{G}}(\mathbf{A}, \mathbf{B} | \mathbf{C})$.

Теорема 2.1 из [3]. При выполнении условия Маркова ациклический ориентированный граф $\mathbf{G}$ порождает те и только те условные независимости, которые отождествляются d-сепарацией.

В замечании к теореме 2.1 из [3] отмечено, что существуют вероятностные распределения P, удовлетворяющие условию Маркова на $\mathbf{G}$, для которых услов-

ные независимости не опознаются d-сепарацией. Далее приведен пример 1, опровергающий результат теоремы 2.1 из [3], но одновременно утверждающий, что множество таких примеров мало по сравнению с теми примерами, для которых теорема выполняется.

В разд. 2.2 работы [3] на основе теоремы 2.1 приведены некорректные утверждения относительно эквивалентности графов по Маркову. В разд. 2.3.1 в [3] дано определение достоверности (faithfulness) и приведена теорема 2.5, которая по содержанию аналогична теореме 2 настоящей статьи.

Теорема 2.5 из [3]**.** Граф **G** и вероятностное распределение P удовлетворяют условию достоверности тогда и только тогда, когда все условные независимости для P определяются с помощью d-сепарации.

В работе [3] утверждается, что доказательство непосредственно вытекает из теоремы 2.1. Очевидно, что теоремы 2.1 и 2.5 противоречат одна другой, поскольку в теореме 2.5 имеется условие достоверности, а в теореме 2.1 его нет, т.е. отсюда следует ошибочность теоремы 2.1.

Покажем, что для дискретных объектов, построенных на ограниченных выборках, теорема 2.1 из [3] и замечание к ней не выполняются для достаточно большого числа контрпримеров.

Пример 1. Рассмотрим граф $(\mathbf{G}, P)$ $X \to Y \to Z$ со следующими переходными вероятностями с шестью параметрами:

$$P(x1) = a,\ P(x2) = 1 - a;\ P(y1|x1) = 1 - (b + c),\ P(y2|x1) = c,\ P(y3|x1) = b;$$

$$P(y1|x2) = 1 - (b + d),\ P(y2|x2) = d,\ P(y3|x2) = b;$$

$$P(z1|y1) = e,\ P(z2|y1) = 1 - e,\ P(z1|y2) = e,\ P(z2|y2) = 1 - e,$$

$$P(z1|y3) = f,\ P(z2|y3) = 1 - f. \qquad (2)$$

Переменные X и Z принимают два значения, переменная Y — три значения. Граф **G** с вероятностным распределением (2) удовлетворяет условию Маркова: выполняется свойство $I_P(X, Z|Y)$. Легко показать, что из (2) вытекает $I_P(X, Z)$, однако $I_{\mathbf{G}}(X, Z)$ не имеет места, т.е. лемма 2.3 из [3] не справедлива. Условие достоверности не выполняется, так как из условия Маркова вытекает $I_P(X, Z|Y)$, а $I_P(X, Z)$ из него не выводится. В соответствии с моделью цепей Маркова имеем

$$P(x1, y1, z1) = a(1 - (b + c))e,\ P(x1, y2, z1) = ace,$$

$$P(x1, y3, z1) = abf,\ P(x1, z1) = a(e + bf - be);$$

$$P(x1, y1, z2) = a(1 - (b + c))(1 - e),\ P(x1, y2, z2) = ac(1 - e),$$

$$P(x1, y3, z2) = ab(1 - f),\ P(x1, z2) = a(1 - e + be - bf);$$

$$P(x2, y1, z1) = (1 - a)(1 - (b + d))e,\ P(x2, y2, z1) = (1 - a)de,$$

$$P(x2, y3, z1) = (1 - a)bf,\ P(x2, z1) = (1 - a)(e + bf - be);$$

$$P(x2, y1, z2) = (1 - a)(1 - (b + d))(1 - e),\ P(x2, y2, z2) = (1 - a)d(1 - e),$$

$$P(x2, y3, z2) = (1 - a)b(1 - f),\ P(x2, z2) = (1 - a)(1 - e + be - bf). \qquad (3)$$

Отсюда следует, что $P(z1) = e + bf - be$, т.е. выполняется независимость $P(X, Z) = P(X)P(Z)$.

Контрпример строится следующим образом. Объект идентифицируется тремя переменными: X определяет форму объекта ($x1$ — квадрат, $x2$ — круг), Y — цифровую информацию ($y1 = 1,\ y2 = 2,\ y3 = 3$), Z — цвет ($z1$ — черный, $z2$ — белый). Определив параметры вероятностного распределения (1) в виде рациональ-

ных чисел $a = c = e = \frac{1}{2}$, $b = d = f = \frac{1}{4}$, получим выборку объема 32, причем количество объектов $m(x, y, z)$ с различными значениями переменных (x, y, z) выражается в виде целых чисел:

$$\begin{aligned}
&m(x1, y1, z1) = 2, && m(x1, y1, z2) = 2,\\
&m(x1, y2, z1) = 4, && m(x1, y2, z2) = 4,\\
&m(x1, y3, z1) = 1, && m(x1, y3, z2) = 3,\\
&m(x2, y1, z1) = 4, && m(x2, y1, z2) = 4,\\
&m(x2, y2, z1) = 2, && m(x2, y2, z2) = 2,\\
&m(x2, y3, z1) = 1, && m(x2, y3, z2) = 3.
\end{aligned} \tag{4}$$

В таком случае можно говорить о конструктивном определении объектов.

Подтверждаем независимость переменных X и Z: $P(x1) = \frac{1}{2}$, $P(x1 | z1) = \frac{1}{2}$, зависимость переменных X и Y: $P(y2) = \frac{3}{8}$, $P(y2 | x1) = \frac{1}{2}$, а также зависимость переменных Y и Z: $P(z1) = \frac{7}{16}$, $P(z1 | y3) = \frac{1}{4}$.

Аналогично при других значениях параметров в (2) получаем примеры на коротких выборках, для которых условие достоверности не выполняется.

Рассмотрим граф $X \leftarrow Y \rightarrow Z$. Вероятностное распределение (2) для этого графа удовлетворяет условию Маркова, поскольку выполняется $I_P(X, Z | Y)$. Поэтому

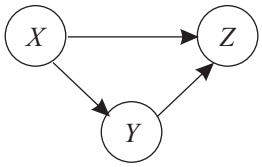
$$P(X, Y, Z) = P(Y)P(X | Y)P(Z | Y). \tag{5}$$

Легко заметить, что (5) совпадает с вероятностным распределением (3), то же справедливо и для графа $X \rightarrow Y \leftarrow Z$ с коллайдером в вершине Y. Вероятностное распределение (2) для этого графа удовлетворяет условию Маркова, поскольку выполняется $I_P(X, Z)$. Поэтому вероятностное распределение

$$P(X, Y, Z) = P(X)P(Z)P(Y | X, Z)$$

этого графа совпадает с вероятностным распределением (3) графа $X \rightarrow Y \rightarrow Z$. В этом случае можно утверждать, что рассмотренные три структуры графа являются эквивалентными. Отметим, что такое понятие эквивалентности не совпадает с определением эквивалентности графов в [3].

Рассмотрим полный ациклический граф с тремя дугами:



(6)

Существует шесть вариантов полных ациклических графов с разными направлениями дуг, два варианта графов имеют циклы. Вероятностное распределение (2) для графа (6) также удовлетворяет условию Маркова, поскольку оно не порождает ни одной условной независимости. Это означает, что не только P (см. (2)), но и любое вероятностное распределение удовлетворяет условию Маркова для каждого варианта полного ациклического графа. Поэтому вероятностное распределение графа (6) $P(X, Y, Z) = P(X)P(Y | X)P(Z | X, Y)$ совпадает с вероятностным распределением (3) графа $X \rightarrow Y \rightarrow Z$.

Наличие дуги между вершинами X и Z в графе (6) не означает, что переменные X и Z не могут быть условно независимыми при некотором заданном множестве переменных. Действительно, для вероятностного распределения (3) вы-

полняется как свойство $I_P(X, Z|Y)$, так и $I_P(X, Z)$. Еще раз убеждаемся, что лемма 2.3 из [3] не справедлива.

Пример 2. Рассмотрим предыдущий граф $X \to Y \to Z$, у которого переходные вероятности определяются восьмеркой параметров:

$$P(x1) = a,\ P(x2) = 1 - a;\ P(y1|x1) = 1 - (b + c),\ P(y2|x1) = c,\ P(y3|x1) = b;$$

$$P(y1|x2) = 1 - (d + e),\ P(y2|x2) = d,\ P(y3|x2) = e;$$

$$P(z1|y1) = f,\ P(z2|y1) = 1 - f,\ P(z1|y2) = j,\ P(z2|y2) = 1 - j,$$

$$P(z1|y3) = h,\ P(z2|y3) = 1 - h.$$

Легко заметить, что условие достоверности в этом примере выполняется, т.е. независимость $I_P(X, Z)$ не имеет места. Присвоив параметрам a, b, c, d, e, f, j, h соответственно рациональные числа $1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9$, получим ограниченную выборку, для которой количество объектов $m(x, y, z)$ с различными значениями переменных (x, y, z) выражается в виде целых чисел.

Сделав перестановку чисел $1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9$, получим $8! = 40\ 320$ различных примеров или вероятностных распределений, для которых выполняется условие достоверности.

Подсчитаем количество примеров, для которых условие достоверности не выполняется. В примере 1 имеется шесть параметров: a, b, c, d, e, f. Из восьми чисел: $1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9$ можно выбрать шесть различных чисел $\mathbf{C}_8^6 = 28$ способами и присвоить их параметрам a, b, c, d, e, f. Для каждой шестерки чисел можно получить $6! = 720$ различных вариантов примеров. В итоге получаем $\mathbf{C}_8^6 6! = 20\ 160$ различных примеров, для которых условие достоверности не выполняется, т.е. ровно в два раза меньше, чем в примере 2.

Пример 3. Рассмотрим граф, у которого переходные вероятности определяются семеркой параметров:

$$P(x1) = a,\ P(x2) = 1 - a;\ P(y1|x1) = 1 - (b + c),\ P(y2|x1) = c,\ P(y3|x1) = b;$$

$$P(y1|x2) = 1 - (d + e),\ P(y2|x2) = d,\ P(y3|x2) = e;$$

$$P(z1|y1) = f,\ P(z2|y1) = 1 - f,\ P(z1|y2) = j,\ P(z2|y2) = 1 - j,$$

$$P(z1|y3) = j,\ P(z2|y3) = 1 - j.$$

Как и в примере 2, условие достоверности в примере 3 выполняется, т.е. имеет место зависимость переменных X и Z. Сделав перестановку семи чисел: $1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8$, получим $7! = 5040$ различных примеров, для которых выполняется условие достоверности. Существует $\mathbf{C}_7^6 = 7$ способов выбора из этих чисел шесть различных чисел и присвоить их параметрам a, b, c, d, e, f. Получаем $\mathbf{C}_7^6 6! = 7!$ различных примеров, для которых условие достоверности не выполняется, и столько же примеров, для которых оно выполняется.

ЗАКЛЮЧЕНИЕ

Известно, что современная математика строится аксиоматически на принципах дедуктивного подхода. Построения и выводы, как правило, выполняются на вещественном непрерывном пространстве (континууме). Континуальное множество в основном состоит из иррациональных чисел, которые являются бесконечными последовательностями десятичных знаков, т.е. формально такого мно-

жества не существует. По свойству континуума счетное и ограниченное множества образуют множество точек меры нуль. Поэтому аналитический аппарат непрерывной математики невозможно адекватно использовать для анализа дискретных объектов и процессов, построенных на ограниченных множествах.

Это стало понятно в 2003 г., когда в результате завершения проекта «Геном человека» выяснилось, что вся генетическая информация о человеке записана в четырехбуквенном алфавите нуклеотидов и ограничена тремя миллиардами букв. Поэтому ограниченный объем генома человека с позиций континуума также имеет меру нуль, что, по меньшей мере, абсурдно. Например, структурные особенности записи генетической информации в ДНК в виде симметрии выводят индуктивно из реальных данных, не опираясь на аппарат непрерывной математики [4, 5].

Для некоторых иррациональных чисел

$$e = 1 + \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{1 \cdot 2 \ldots n} + \ldots, \quad \pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \ldots\right)$$

существуют конечные правила определения десятичных знаков. То же самое можно сказать о любом рациональном числе и некоторых иррациональных, поскольку таких конечных правил не более чем счетное число. Таким образом, почти весь континуум состоит из чисел, которые конструктивно не определяются. Но даже при наличии алгоритма бесконечное количество знаков — в некотором роде фикция, ибо процесс вычислений никогда не заканчивается. С позиции классического анализа декларируется, что бесконечность состоялась — безостановочная работа проведена до конца, и в результате получено иррациональное число. В настоящее время в непрерывной математике накоплено много различного рода контрпримеров и парадоксов, ставящих под сомнение использование такого искусственного бесконечного множества [6].

Однако континуум содержит любое ограниченное дискретное множество и на таком искусственном множестве можно получать «посторонние» результаты, как это показано в настоящей работе, которые не выполняются для дискретных (реальных) объектов, построенных на ограниченных выборках. Результаты в [1–3] не дают правильного представления о свойствах сепарационного подхода. По сути эти результаты выражают свойства непрерывных моделей на вещественном пространстве. Непрерывный случай не представляет практического интереса, поскольку основное назначение процедур сепарации заключается в определении структуры байесовской сети из ограниченного набора данных и бесконечное континуальное множество к этому отношения не имеет. Таким образом, результаты в [1–3] и в других монографиях требуют кардинального пересмотра с точки зрения изложения свойств сепарационного подхода для дискретных объектов.

СПИСОК ЛИТЕРАТУРЫ

1. Pearl J. CAUSALITY: models, reasoning, and inference. — Cambridge Univ. press, 2000. — 526 p.
2. Spirtes P., Glymour C., Scheines R. Causation, prediction, and search (2nd Ed.). — New York: MIT press, 2001. — 544 p.
3. Neapolitan R.E. Learning Bayesian networks. — NJ, Upper Saddle River: Prentice Hall, 2004. — 696 p.
4. Сергиенко И.В., Гупал А.М., Вагис А.А. Правила симметрии в записи генетической информации в ДНК // Кибернетика и системный анализ. — 2011. — № 3. — С. 88–94.
5. Сергиенко И.В., Гупал А.М., Вагис А.А. Симметрия и свойства записи информации в ДНК // Докл. НАНУ — 2011. — **439**. — № 1 — С. 30–32.
6. Босс В. Лекции по математике. — М.: Книжный дом «ЛИБРОКОМ», 2009. — Т. 12: Контпримеры и парадоксы. Учебное пособие. — 216 с.

Поступила 10.11.2011