УДК 007:681.3.00

А.С. БАЛАБАНОВ

МИНИМАЛЬНЫЕ СЕПАРАТОРЫ В СТРУКТУРАХ ЗАВИСИМОСТЕЙ. СВОЙСТВА И ИДЕНТИФИКАЦИЯ

Ключевые слова: *ациклические орграфы, АОГ-модель вероятностных зависимостей, условная независимость, байесовские сети, критерий d-сепарации, коллайдер, цепь, локально-минимальные сепараторы, идентификация ребер, подбор сепараторов, необходимые требования к членам локально-минимальных сепараторов.*

ВВЕДЕНИЕ

В настоящей работе рассматриваются вероятностные модели зависимостей, структурированные ациклическими орграфами (АОГ-модели). Различают три основные разновидности АОГ-моделей: байесовские сети, гауссовы сети и гибридные сети. В байесовских сетях переменные — номинальные (дискретные), в гауссовых — действительные (зависимости — линейные). В гибридных сетях используются оба типа переменных. Результаты статьи охватывают все разновидности АОГ-моделей, так как базируются только на общих свойствах этих моделей — ацикличности орграфа и критерии *d*-сепарации.

АОГ-модели — наглядный и компактный язык представления систем зависимостей [1–7]. Они являются эффективным аппаратом отображения связей (вплоть до причинно-следственных отношений) и инструментом вероятностных рассуждений от свидетельств (в экспертных системах). Эти свойства делают АОГ-модели средством решения разнообразных задач: медицинская и техническая диагностика, распознавание речи, прогнозирование последствий решений и действий, анализ данных в эконометрике, социометрике, микробиологии и т.д. [3–5, 8]. Для решения разных познавательных, исследовательских и прогнозно-аналитических задач необходимо выводить структуру модели на основе выборки данных наблюдений за моделируемой системой. Известно, что задача вывода АОГ-модели из данных в общем случае — NP-сложная [9]. Вычислительные сложности становятся серьезной практической проблемой, когда число переменных модели достигает нескольких десятков. Помимо того, существует проблема надежности вывода модели. Поэтому на практике часто приходится полагаться на априорные знания или принимать жесткие предположения и (или) вынужденные ограничения, тем самым снижая ценность вывода.

Проблема поиска и подбора сепараторов в моделях зависимостей возникает в различных видах задач, и прежде всего — при выводе АОГ-модели из данных методами так называемого constraint-based («сепарационного») подхода [2, 3, 5, 7]. Поиск сепараторов в сложных структурах — комбинаторная задача. Важно как можно раньше распознать ситуацию, когда предполагаемого сепаратора не существует, и прекратить бесперспективный поиск. От состава сепаратора зависит объем вычислений и надежность тестов независимости, поэтому желательно находить минимальные сепараторы. Минимизация сепараторов также имеет значение в задачах рассуждений с использованием модели (при выводе следствий из свидетельств), так как информация распространяется по структуре модели через сепараторы. Сформулировав требования к элементам сепараторов,

можно отсеять некоторые кандидаты в сепаратор и тем самым упростить задачу. В работе показаны новые возможности в поиске и подборе сепараторов в АОГ-моделях, что должно обеспечить повышение эффективности решения соответствующих задач вывода. Для развернутой характеристики проблемы требуется очертить теоретические основы АОГ-моделей зависимостей.

ОСНОВЫ ТЕОРИИ АОГ-МОДЕЛЕЙ ЗАВИСИМОСТЕЙ. СЕПАРАЦИЯ. ПРОБЛЕМА

Вероятностные модели зависимостей на основе графов — актуальная тема современных исследований на стыке многомерного статистического анализа, теории графов, теории информации и искусственного интеллекта. Вероятностные графовые модели зависимостей — строгий язык представления знаний в условиях неполноты и неопределенности. Наиболее популярны АОГ-модели. К достоинствам АОГ-моделей относятся наглядность, способность отображать причинно-следственные связи, компактное (по числу параметров) представление систем зависимостей и вычислительная эффективность вероятностного вывода от свидетельств [1–4]. Вероятностные графовые модели зависимостей можно идентифицировать индуктивно, на основе статистических данных (опираясь на несколько методологических постулатов). Таким образом, это один из подходов к открытию знаний в базах данных.

Зафиксируем элементарные графовые понятия. Если в орграфе G есть дуга $x \to y$, то вершина x называется «родителем» вершины y, а вершина y — ребенком вершины x. Ребро — это дуга, ориентация которой неизвестна или игнорируется. Вершины x и y называются смежными, если они соединены ребром, что обозначается как x — y. Отсутствие ребра будем обозначать как $\neg (x$ — $y)$. Путь в орграфе — последовательность смежных ребер (т.е. дуг любой ориентации), без повторения вершин. Как исключение, первая и последняя вершины пути могут совпадать, и тогда этот путь называют циклом. Орпуть (строго ориентированный путь) — путь, на котором все ребра ориентированы в направлении одного и того же конца пути. Ациклический ориентированный граф (АОГ) — орграф, в котором нет ориентированных циклов. Ввиду ацикличности любой АОГ содержит корни (корень). Далее под орграфом будем подразумевать АОГ. Если в графе есть орпуть $x \to \cdots \to y$, то вершина x называется предком вершины y.

АОГ-модель — это модель, структура которой — орграф без орциклов (и петель), причем каждой переменной соответствует вершина графа. АОГ-модель определяется как (G, θ), где G — АОГ, а θ — совокупность локально заданных параметров. Более известны два вида АОГ-моделей: байесовские сети и гауссовы сети. (Заметим, что наименование «байесовские сети» широко используется по отношению ко всем АОГ-моделям.) Не будем различать разновидности АОГ-моделей, так как эксплуатируем только графовые свойства моделей (не параметрические). (Линейность модели облегчает анализ и предотвращает некоторые осложнения.) Предлагаемые средства более важны для байесовских сетей, поскольку для них сложнее вычислять статистики для тестов. Ввиду взаимно однозначного соответствия термины «переменная» (модели) и «вершина» (графа) в литературе употребляются как взаимозаменяемые.

Определение 1. Коллайдером (коллизором) в графе называется фрагмент вида $x \to y \leftarrow z$. Если коллайдер $x \to y \leftarrow z$ является частью пути π в орграфе, то y называется коллайдерной вершиной на пути π. (При это вершина y может быть неколлайдерной на некотором другом пути.)

Бесколлайдерный (бесколлизорный) путь, или цепь, в орграфе — путь, не содержащий ни одного коллайдера.

Отношение между структурой модели и фактами условной независимости, представленными в АОГ-модели, формализовано с помощью критерия d-сепарации [1, 3, 4, 6]. Перед формулировкой определения d-сепарации целесообразно дать пояснения. В оригинале определения используется простая конструкция «given **S**» (где **S** — множество вершин). Буквальный перевод, очевидно, не годится. По смыслу можно было бы сказать «при блокировании **S**» или «при фиксации состояния множества вершин **S**». Однако слово «блокирование» принято употреблять по отношению к путям (не вершинам). Второй вариант перевода — громоздкий, к тому же потребует определить, что такое «состояние вершины» и «фиксация». (Впрочем, специалисты могут мыслить именно такими понятиями.) Поэтому предлагаем использовать слово «кондиционирование» (т.е. введение в условие). Этой операции на графе соответствует одноименная операция над данными (или над распределением вероятностей). Когда по контексту роль множества **S** понятна, можно обойтись без уточняющего термина, а просто сказать «с помощью множества вершин **S**».

Определение 2 (d-сепарация). Путь π в АОГ-модели называют d-закрытым (d-блокированным) с помощью (кондиционирования) множества вершин **S**, если и только если выполняется хотя бы одно из следующих условий:

1) существует вершина x, $x \in \mathbf{S}$, $x \in \pi$, причем на пути π есть дуга $x \rightarrow$ или $\leftarrow x$;

2) на пути π есть хотя бы один коллайдер $\rightarrow y \leftarrow$, причем $y \notin \mathbf{S}$ и вершина y не является предком никакой вершины в **S**, т.е. не существует никакой $z \in \mathbf{S}$, такой, что есть $y \rightarrow \cdots \rightarrow z$.

Множество вершин **S** d-сепарирует вершины x и y, если и только если все пути между x и y являются d-закрытыми с помощью множества вершин **S**. Будем обозначать такую d-сепарацию предикатом $\mathrm{Ds}(x \perp \mathbf{S} \perp y)$. Когда $\mathbf{S} = \varnothing$, будем писать $\mathrm{Ds}(x \perp \perp y)$. Если хотя бы один путь между x и y не является d-закрытым, то говорят, что вершины x и y d-соединены. Это обозначается $\neg\, \mathrm{Ds}(x \perp \mathbf{S} \perp y)$.

Определение 3 (сепаратор). Если в АОГ предикат $\mathrm{Ds}(x \perp \mathbf{S} \perp y)$ истинен, то множество вершин **S** называют (графовым) сепаратором для пары (x, y).

Понятие сепаратора естественным образом расширяется на случай $\mathrm{Ds}(\mathbf{X} \perp \mathbf{S} \perp \mathbf{Y})$, где **X** и **Y** — множества вершин. Очевидно, в любой графовой модели (даже более общей, чем АОГ) из существования ребра x — y следует отсутствие сепаратора для (x, y). Обратная импликация верна только для АОГ, т.е. в моделях без скрытых переменных.

Цель аналитика при решении многих познавательных и исследовательских проблем — вывести структуру модели из статистических данных наблюдений за моделируемой системой. Эта задача сводится к идентификации всех дуг графа модели. Наличие или отсутствие ребер следует верифицировать на основе соответствующих статистических свойств эмпирических данных, а именно, фактов условной независимости. Условную независимость переменной x от переменной y при кондиционировании множества переменных **S** обозначим $\mathrm{Pr}(x \perp \mathbf{S} \perp y)$. При этом подразумевается, что $x, y \notin \mathbf{S}$. Для дискретных переменных эта условная независимость означает $p(y|\mathbf{S}, x) = p(y|\mathbf{S})$. В линейных моделях условная независимость проявляется как нулевое (в конечной выборке — незначимо отличающееся от нуля) значение коэффициента частной корреляции. Безусловную независимость, например $\mathrm{Pr}(x \perp \varnothing \perp y)$, запишем в форме $\mathrm{Pr}(x \perp\perp y)$. Факт безусловной зависимости обозначим $\neg\, \mathrm{Pr}(x \perp\perp y)$. (Напомним, что в конечной выборке не всякие зависимости статистически значимы.)

Большинство известных методов вывода структуры модели из статистических данных относятся либо к оптимизационному, либо к «сепарационному» (constraint-based) подходу [2, 3, 5, 7]. Последний основан на выявлении фактов условной независимости.

Известно [6], что из d-сепарации следует истинность соответствующего утверждения условной независимости, т.е.

$$\mathrm{Ds}\,(x \perp \mathbf{S} \perp y) \Rightarrow \mathrm{Pr}\,(x \perp \mathbf{S} \perp y). \tag{1}$$

Таким образом, критерий d-сепарации обеспечивает считывание утверждений условной независимости из графа АОГ-модели. Однако для вывода модели из данных нужны правила обратного вида.

Основным методологическим постулатом сепарационных методов является предположение необманчивости (faithfulness) распределения вероятностей АОГ-модели [2, 3, 7, 10], которое можно выразить как

$$\mathrm{Pr}\,(x \perp \mathbf{S} \perp y) \Rightarrow \mathrm{Ds}\,(x \perp \mathbf{S} \perp y), \tag{2}$$

т.е. как обратную импликацию по сравнению с (1). Сопоставляя (1) и (2), получаем

$$\mathrm{Ds}\,(x \perp \mathbf{S} \perp y) \Leftrightarrow \mathrm{Pr}\,(x \perp \mathbf{S} \perp y). \tag{3}$$

Таким образом, выполнение (2) обеспечивает структурно-поведенческий изоморфизм модели.

Сопоставляя свойство АОГ-моделей $(x \text{ — } y) \Leftrightarrow \neg\,\exists\,\mathbf{S}\colon \mathrm{Ds}\,(x \perp \mathbf{S} \perp y)$ и эквивалентность (3), получаем

$$x \text{ — } y \Leftrightarrow \neg\,\exists\,\mathbf{S}\colon \mathrm{Pr}\,(x \perp \mathbf{S} \perp y). \tag{4}$$

Принцип (4) лежит в основе большинства сепарационных методов, в частности алгоритма PC [2, 3, 5]. Заметим, что (4) опирается на ослабленную форму предположения необманчивости (которая характеризует только ребра, а не любые пути), но этого достаточно для построения результативного алгоритма идентификации всех ребер модели. Однако даже такая ослабленная форма предположения необманчивости может изредка нарушаться, что создает риск ошибки.

Определение 4[11, 12]. Локально-минимальным сепаратором для пары вершин (x, y) в АОГ называется такой сепаратор $\mathbf{S}$, что после удаления из $\mathbf{S}$ любого его члена (элемента) он перестает быть сепаратором для (x, y). Формально это записывается как

$$\mathrm{Ds}\,(x \perp \mathbf{S} \perp y)\ \ \&\ \ \forall\, z \in \mathbf{S}\colon \neg\,\mathrm{Ds}\,(x \perp \mathbf{S} \setminus \{z\} \perp y).$$

Определение 5 [12]. Назовем сепаратор $\mathbf{S}^*$ для пары вершин (x, y) в АОГ минимальным сепаратором, если для (x, y) не существует сепаратора меньшей кардинальности, т.е. для всех других сепараторов $\mathbf{S}$ для пары (x, y) верно $|\mathbf{S}| \geq |\mathbf{S}^*|$.

Обозначим минимальный и локально-минимальный сепаратор для пары вершин (x, y) соответственно $S_{\min}(x, y)$ и $S_{\mathrm{lom}}(x, y)$.

При анализе свойств сепарации в АОГ-моделях будем опираться только на критерий d-сепарации и запрет орциклов. В этом анализе нельзя воспользоваться известными методами и результатами теории графов (и транспортных задач) ввиду следующих особенностей:

1) зависимости блокируются манипуляцией вершин (переменных), а не ребер;

2) свойства d-сепарации существенно отличаются от транспортных аналогов;

3) обычно в момент поиска сепаратора структура графа еще неизвестна (известны только некоторые факты и ограничения).

В АОГ-моделях ребра соответствуют ассоциациям (зависимостям). Блокировать сами ребра аналитик не может. Никакие операции вроде поиска разреза графа недоступны. При анализе данных (распределений вероятностей) доступны только операции с переменными. Следует отметить и более тонкое отличие d-сепарации от блокирования потоков в транспортных сетях. Согласно определению 2 (условие 2) возможно разблокирование пути «сбоку», «издали». Механический аналог этого явления неизвестен (например, аналогия с семафором — неточна). В [13] это описано как провоцирование зависимости. Можно интуитивно представить, что в моделях зависимостей пути представляют «потоки влияния» либо «цены доставки», однако эти суррогатные понятия не будут подчиняться ни ограничениям аддитивности цены, ни баланса «дебет-кредит». (Как исключение — в линейных моделях выполняется аддитивность влияний через параллельные пути.)

Определение 6 (эмпирический сепаратор). Если в (выборочном) распределении вероятностей выполняется $\Pr(x \perp \mathbf{S} \perp y)$, то множество переменных $\mathbf{S}$ называют эмпирическим, или статистическим, сепаратором для пары (x, y).

Эмпирический сепаратор для пары (x, y) может не совпадать ни с одним графовым сепаратором для (x, y) в модели, что означает нарушение предположения необманчивости. Самый худший случай — когда существует ребро x — y и, тем не менее, существует эмпирический сепаратор для (x, y). Тогда принцип (4) допускает сбой.

Одним из недостатков методов сепарационного подхода является ненадежность идентификации ребер модели (риск ошибок). Этот риск в основном объясняется ненадежностью результатов тестирования утверждений условной независимости, когда аналитик располагает выборкой данных недостаточно большого объема. Другая проблема — неудовлетворительная вычислительная эффективность поиска сепараторов. Вычислительная эффективность определяется количеством выполняемых тестов и их сложностью (т.е. объемом вычисления необходимых статистик.)

Задача поиска сепараторов при выводе структуры АОГ-модели из данных является переборной, и чем больше имеется кандидатов в сепаратор, тем сложнее поиск. Для оптимизации поиска сепараторов алгоритм PC ограничивает множество кандидатов в члены сепаратора для (x, y) множеством вершин, которые считаются смежными с x или y на данном этапе вывода модели. Такая тактика обеспечивает корректный вывод при условии выполнения (4). Однако она несовершенна в том, что алгоритм: 1) не всегда находит минимальные сепараторы; 2) продолжает искать сепаратор до полного исчерпания вариантов даже тогда, когда сепаратора не существует. Вследствие этого алгоритм рискует допустить ошибку при тестировании условной независимости с большой кардинальностью условия.

Порядок (ранг) теста условной независимости $\Pr(x \perp \mathbf{S} \perp y)$ определяется кардинальностью условия $\mathbf{S}$. С ростом кардинальности $\mathbf{S}$ происходит, по существу, дробление выборки данных. Фактор дробления выборки (в дискретных моделях) имеет порядок величины $||x||\cdot||y||\cdot||\mathbf{S}||$, где $||\mathbf{S}|| = \Pi_i ||z_i||$, $z_i \in \mathbf{S}$, $||x||$ — значность (арность) переменной x. Этот же фактор оценивает объем вычисления статистик, необходимых для теста.

Таким образом, для надежного и эффективного восстановления «истинной» структуры модели необходимо по возможности обходиться тестами низкого порядка. Отсюда вытекает значение и важность задачи поиска минимальных сепараторов и задача быстрого распознавания ситуации, когда сепаратора не существует.

Ключевая идея предлагаемого анализа — использовать знания об одних сепараторах и зависимостях для характеристики других сепараторов и связей. Естественно, в качестве признаков следует использовать менее сложные конструкции (условия), чем в резолюционной части утверждений и правил. Некоторые возможности в этом направлении (на базе безусловных зависимостей, с использованием понятия генотипов переменных) показаны в [11]. Аппарат генотипов переменных является компактным способом представления множества безусловных (не)зависимостей. В [12] получены новые результаты; в данной работе предлагается их развитие.

СВОЙСТВА ЛОКАЛЬНО-МИНИМАЛЬНЫХ СЕПАРАТОРОВ (И ИХ ЭЛЕМЕНТОВ) В АОГ-МОДЕЛЯХ

Очевидно, что каждый минимальный сепаратор является также и локально-минимальным. Значит, все необходимые требования к членам локально-минимальных сепараторов полностью распространяются на членов минимальных сепараторов. Когда локально-минимальный сепаратор единственный, он минимален. В одном и том же АОГ может быть несколько минимальных сепараторов для пары вершин (x, y). В то же время единственный $S_{\min}(x, y)$ может не иметь ни одного общего члена с некоторым $S_{\text{lom}}(x, y)$. Известно [14], что никакое подмножество множества $S_{\text{lom}}(x, y)$ не может быть сепаратором для пары вершин (x, y). Таким образом, локально-минимальные сепараторы являются «несжимаемыми».

Начнем с констатации элементарных свойств.

Факт 1. Если в графе верно $\text{Ds}(x \perp\!\!\!\perp y)$, то для пары вершин (x, y) имеется только один локально-минимальный сепаратор, а именно — пустой.

Факт 2.Каждый член каждого локально-минимального сепаратора **S** является неколлайдерной вершиной на том пути (путях), который он блокирует, причем этот путь (как минимум один из путей) не блокируется никаким другим членом **S**.

Факт 3. Каждый член x каждого локально-минимального сепаратора имеет не менее одной исходящей дуги $x \to$ и не менее двух безусловно зависимых вершин (т.е. d-соединенных с x).

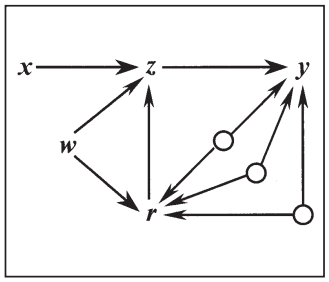

Рис. 1

Эти факты следуют прямо из определений 2 и 4. Заметим, что факт 2 нельзя «прямолинейно» усилить. А именно, нельзя потребовать, чтобы для каждой $z \in S_{\text{lom}}(x, y)$ существовал путь между вершинами x и y через z, такой, что одна из двух частей этого пути была бы цепью. (Такое ошибочное требование входит в формулировку утверждения, представленного в [11].) Пример невыполнения этого требования иллюстрируется на рис. 1. Для данной модели множество $\{z, r, w\}$ является минимальным сепаратором для (x, y). Легко видеть, что вершина w не лежит ни на какой цепи между вершинами x и y. Более того, нет ни одного пути между вершинами x и y через w, частью которого была бы цепь $x — \cdots — w$ или цепь $w — \cdots — y$.

Факт 4(стыковка). Если в графе есть хотя бы одна цепь $z — \cdots — x$ и хотя бы один орпуть $x \to \cdots \to y$, то существует по крайней мере одна цепь между z и y.

(При этом допустимо пересечение названных цепи и орпути.)

Из факта 4 логически вытекает следующее.

Факт 5(правило запрета дуги):

$$\exists\, w: \mathrm{Ds}\,(w \perp\!\!\!\perp y) \,\&\, \neg\, \mathrm{Ds}\,(w \perp\!\!\!\perp x) \Rightarrow \neg\,(x \to y); \tag{5}$$

$$\exists\, z: \mathrm{Ds}\,(z \perp\!\!\!\perp x) \,\&\, \neg\, \mathrm{Ds}\,(z \perp\!\!\!\perp y) \Rightarrow \neg\,(y \to x). \tag{5a}$$

Утверждение 1 (правило непоглощения). Если в АОГ имеем $\neg\, \mathrm{Ds}\,(x \perp\!\!\!\perp y)$ и существуют такие вершины z и w, что $\mathrm{Ds}\,(z \perp\!\!\!\perp y)$, $\neg\, \mathrm{Ds}\,(z \perp\!\!\!\perp x)$, $\mathrm{Ds}\,(w \perp\!\!\!\perp x)$ и $\neg\, \mathrm{Ds}\,(w \perp\!\!\!\perp y)$, то невозможно ребро x — y, т.е. существует сепаратор для пары вершин (x, y).

Понятно, что когда условия выполнены, все цепи между вершинами x и y имеют вид $x \leftarrow \cdots \to y$. Другая форма этих же, по сути, правил дана в [11], где они выражены через аппарат генотипов переменных.

Утверждение 2. В составе каждого непустого локально-минимального сепаратора для пары вершин (x, y) есть, как минимум, одна вершина, которая лежит на некоторой цепи между вершинами x и y.

Эта формулировка эквивалентна утверждению, доказанному в [12]. Утверждение 2 нельзя усилить, в том смысле, что возможны случаи, когда все члены локально-минимального сепаратора, за исключением одного, не лежат ни на каких цепях между вершинами x и y. Иллюстрацией служит рис. 1.

Утверждение 2а. В составе каждого непустого локально-минимального сепаратора для пары вершин (x, y) есть, как минимум, одна вершина z такая, что $\neg\, \mathrm{Ds}\,(z \perp\!\!\!\perp x) \,\&\, \neg\, \mathrm{Ds}\,(z \perp\!\!\!\perp y)$.

Факт 6 (правило единственного общего близкого; single common covariate). Если в АОГ нет ребра x — y и $\neg\, \mathrm{Ds}\,(x \perp\!\!\!\perp y)$ и есть только одна вершина z, d-соединенная с обеими вершинами x и y, то вершина z входит в состав всех сепараторов для пары вершин (x, y). Если при этих условиях нет ни одной вершины w ($w \neq x, y$), d-соединенной с вершиной z, то $\{z\}$ является единственным локально-минимальным сепаратором для (x, y).

Факт 7. Пусть в орграфе нет ребра x — y, но есть путь x — z — y. Тогда если этот путь имеет вид $x \to z \leftarrow y$, то ни вершина z, ни какой-либо ее потомок не входит в состав ни одного сепаратора для пары вершин (x, y). В трех других случаях ориентации пары ребер x — z — y вершина z является обязательным членом любого сепаратора для пары вершин (x, y).

Неколлайдерная вершина z на пути x — z — y является «стержнем» (pivot) сепаратора для (x, y).

Факт 8 (правило множества изолированных общих близких; set of isolated common covariates). Пусть в АОГ верно $\neg\, \mathrm{Ds}\,(x \perp\!\!\!\perp y)$ и $\mathbf{R}$ — множество всех вершин, зависимых одновременно от x и y, т.е. $\mathbf{R} = \{r \mid \neg\, \mathrm{Ds}\,(r \perp\!\!\!\perp x) \,\&\, \neg\, \mathrm{Ds}\,(r \perp\!\!\!\perp y)\}$, причем $|\mathbf{R}| \geq 2$, и для каждой пары $r, z \in \mathbf{R}$ верно $\mathrm{Ds}\,(r \perp\!\!\!\perp z)$. Тогда либо существует ребро x — y, либо множество $\mathbf{R}$ — единственный локально-минимальный сепаратор для пары (x, y).

Факт 8 следует из фактов 4, 7. Заметим, что (в условиях факта 8) все вершины множества $\mathbf{R}$ являются корневыми, а вершины x и y не имеют ни одного ребенка, если не брать во внимание возможного ребра x — y. Возможны три случая: 1) все $r \in \mathbf{R}$ являются общими родителями для x и y; 2) подмножество вершин множества $\mathbf{R}$ — общие родители для x и y, а остальные члены $\mathbf{R}$ — родители только вершины x; 3) все вершины $r \in \mathbf{R}$ являются родителями только вершины x. (Симметричные варианты не называем.) В случае 1) ребро x — y возможно. В случаях 2) и 3) обязательно есть дуга $x \to y$. В случае 3) верно $\mathrm{Ds}\,(\mathbf{R} \perp\!\!\!\perp x \perp\!\!\!\perp y)$.

Факт 9. В произвольном АОГ сепаратором для пары несмежных вершин (x, y) является множество всех родителей вершины x или множество всех родителей вершины y, или каждое из этих множеств. Также сепаратором для пары (x, y) является объединение всех родителей обеих вершин x и y.

Доказательство. Очевидно, что не могут существовать одновременно орпуть $x \to \cdots \to y$ и орпуть $y \to \cdots \to x$. Пусть (технически) нет ни одного орпути $y \to \cdots \to x$. Тогда сепаратором для (x, y) будет множество родителей вершины y. Предположим противное. Тогда между вершинами x и y существует некоторый путь π, открытый при кондиционировании родителей вершины y. Следовательно, крайней дугой пути π должна быть $y \to$. Значит, если путь π — бесколлайдерный (т.е. цепь), то он имеет вид $y \to \cdots \to x$. Но это противоречит техническому допущению. Если путь π — коллайдерный, рассмотрим ближайший к вершине y коллайдер $\to z \leftarrow$. Ясно, что есть орпуть вида $y \to \cdots \to z$. Так как этот коллайдер должен быть открыт при кондиционировании родителей вершины y, должен существовать орпуть вида $z \to \cdots \to y$. Получаем орцикл. Итак, предположение от противного было неверным. Следовательно, если нет ни одного орпути между вершинами x и y (ни в каком направлении), то сепаратором для (x, y) является и множество всех родителей вершины x, и множество всех родителей вершины y. Остальное доказать легко.

Отметим, что бывают случаи, когда сепаратор для пары вершин (x, y) существует, но никакое подмножество множества вершин, смежных с x, в том числе все это множество, не является сепаратором для (x, y). Понятно, что в таком случае сепаратор можно составить из вершин, смежных с y.

Из факта 9 следует: если в АОГ не существует ребра x — y, то сепаратор для пары (x, y) существует.

Утверждение 3. Если в АОГ вершина w входит в состав некоторого сепаратора для пары вершин (x, y) и существуют вершины q, z такие, что верно $\mathrm{Ds}(w \perp q \perp x)$ и $\mathrm{Ds}(w \perp z \perp y)$, то существует сепаратор для пары вершин (x, y), который не содержит вершину w.

Действительно, таким сепаратором (возможно, не минимальным) является объединение всех родителей вершин x и y. Утверждение 3 легко обобщить следующим образом.

Утверждение 4.Если в АОГ вершина w входит в состав некоторого сепаратора для пары вершин (x, y) и существуют множества вершин $\mathbf{R}$, $\mathbf{S}$ такие, что верно $\mathrm{Ds}(w \perp \mathbf{R} \perp x)$ и $\mathrm{Ds}(w \perp \mathbf{S} \perp y)$, то существует сепаратор для пары вершин (x, y), который не содержит вершину w.

Утверждения 3, 4 и факт 9 оправдывают тактику алгоритма PC, который при поиске сепаратора для (x, y) отбрасывает все вершины, не смежные с x и y. Такая тактика целесообразна с точки зрения быстрого сокращения множества кандидатов в члены сепаратора, но в то же время она часто ведет к потере минимальных сепараторов. Это можно проиллюстрировать на примере модели, изображенной на рис. 2. Алгоритм PC найдет факты $\mathrm{Ds}(w \perp q \perp x)$ и $\mathrm{Ds}(w \perp z \perp y)$ и поэтому исключит вершину w из рассмотрения при поиске сепаратора для (x, y). Значит, в дальнейшем алгоритм уже не сможет найти минимальный сепаратор $\{q, w, z\}$. (Заметим, этот сепаратор состоит из вершин, не смежных с x и y.) Алгоритм PC также найдет факты $\mathrm{Ds}(x \perp q \perp \{g, h\})$ и $\mathrm{Ds}(y \perp z \perp \{r, t\})$. Затем он с помощью сепаратора, состоящего из двух неименованных вершин, показанных на рисунке кружками, исключит

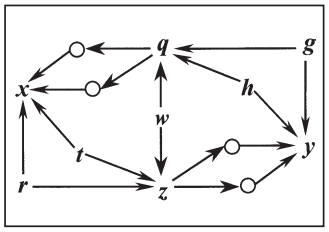

Рис. 2

вершину q из числа смежных с x и, симметрично, исключит вершину z из числа смежных с y. В результате алгоритм уже не сможет найти сепараторы $\{g, h, z\}$ и $\{t, r, q\}$. Следовательно, ни один минимальный сепаратор не будет найден (будет найден сепаратор из четырех переменных).

Легко построить примеры, где сепараторы, найденные алгоритмом РС, будут намного сложнее минимальных. Далее излагаются главные результаты.

Утверждение 5 (базовая теорема о членах локально-минимального сепаратора). Если в АОГ вершина z входит в состав некоторого локально-минимального сепаратора для пары вершин (x, y), то:

а) существует некоторый ориентированный путь от вершины z до вершины x, не проходящий через y, или существует некоторый ориентированный путь от вершины z до вершины y, не проходящий через x;

б) если не существует ни одной цепи между вершинами x и z, которая не проходит через y, то существуют (по меньшей мере) две некоторые цепи λ_1 и λ_2 между вершинами z и y, которые заканчиваются дугами $\rightarrow y$ и не проходят через x.

Заметим, что цепи λ_1 и λ_2 могут взаимно налагаться и пересекаться, но их отрезки, прилегающие к вершине z, не совпадают. Кроме того, ориентированный путь, названный в п. а) теоремы, может совпадать с одной из цепей λ_1 или λ_2, упомянутых в п. б) теоремы.

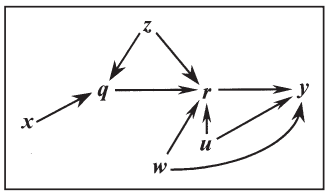

Рис. 3

Доказательство базовой теоремы дано в [12]. Формулировка более раннего варианта теоремы из [11] неточна, что уже отмечалось выше (после факта 3). Базовая теорема иллюстрируется на рис. 3, где локально-минимальными сепараторами для (x, y) являются $\{q, z\}$ и $\{r, u, w\}$. При этом не существует ни одной цепи между вершиной x и вершинами z, u, w (которые являются членами соответствующих $S_{\text{lom}}(x, y)$). В то же время из каждой вершины z, u, w есть по два орпути в y, которые не проходят через x. Причем орпути $z \rightarrow q \rightarrow r \rightarrow y$ и $z \rightarrow r \rightarrow y$ взаимно налагаются. Обращаем внимание, что, несмотря на то, что все цепи между вершинами z и y проходят через вершину r, минимальный сепаратор для (x, y) включает вершину z, а не вершину r. Теперь обратимся к модели на рис. 1 и убедимся, что $r \in S_{\min}(x, y)$. При этом вершина r соединена с вершиной y одним орпутем $r \rightarrow z \rightarrow y$ и еще четырьмя цепями.

Из базовой теоремы немедленно вытекает следующее.

Факт 10. Если верно $\text{Ds}(w \perp\!\!\!\perp x)$ & $\text{Ds}(w \perp\!\!\!\perp y)$, то $w \notin S_{\text{lom}}(x, y)$.

Факт 11 (правило «чужого гена»). Если в орграфе для заданной вершины z существует такая вершина w, что $\neg\, \text{Ds}(w \perp\!\!\!\perp z)$, $\text{Ds}(w \perp\!\!\!\perp x)$, $\text{Ds}(w \perp\!\!\!\perp y)$, то вершина z не входит в состав никакого локально-минимального сепаратора для пары вершин (x, y).

(Заметим, что это правило работает также и в случае существования x — y.) Факт 11 следует из п. а) базовой теоремы и факта 4. Эквивалентная формулировка этого правила была выражена в аппарате генотипов переменных [11].

Факт 12 (правило «изолятора»). Пусть в АОГ имеем $\neg\, \text{Ds}(x \perp\!\!\!\perp y)$ и для вершин z и r верно $\neg\, \text{Ds}(z \perp\!\!\!\perp x)$, $\text{Ds}(z \perp\!\!\!\perp y)$, $\text{Ds}(r \perp\!\!\!\perp x)$, $\neg\, \text{Ds}(r \perp\!\!\!\perp y)$. Тогда если существует такая вершина w $(w \neq x,\ w \neq y)$, что $\neg\, \text{Ds}(w \perp\!\!\!\perp z)$ и $\neg\, \text{Ds}(w \perp\!\!\!\perp r)$, то сепаратор для пары вершин (x, y) существует, а вершина w не входит в состав никакого локально-минимального сепаратора для пары (x, y).

Доказательство. Вывод о существовании сепаратора повторяет утверждение 1 (правило непоглощения). Для доказательства остального предположим

противное, т.е. что $w \in S_{\text{lom}}(x, y)$. Тогда, согласно пункту а) базовой теоремы, должен существовать орпуть $w \to \cdots \to x$ или орпуть $w \to \cdots \to y$. В первом случае ввиду $\neg \operatorname{Ds}(w \perp\!\!\!\perp r)$ и факта 4 получим $\neg \operatorname{Ds}(r \perp\!\!\!\perp x)$, что противоречит условию. Во втором случае аналогично получаем $\neg \operatorname{Ds}(z \perp\!\!\!\perp y)$, т.е. противоречие условию. (По аналогии с фактом 11, можно сказать, что вершина w изолирует два «получужих гена».)

Продолжением и уточнением базовой теоремы служит следующее утверждение.

Утверждение 6. Если верно $z \in S_{\text{lom}}(x, y) = \mathbf{S}$ и $\operatorname{Ds}(x \perp\!\!\!\perp z)$, то:

а) существует некоторая цепь между вершинами x и y, которая не проходит через z;

б) все цепи между вершинами x и y, а также все цепи между вершинами z и y заканчиваются дугами $\to y$;

в) среди всех путей между вершинами x и y, которые проходят через z, есть некоторый путь π, на котором все коллайдеры открыты при кондиционировании $\mathbf{S} \setminus \{z\}$.

Доказательство. Пункт а) следует из утверждения 2. Переходим к п. б). Если бы существовала цепь между вершинами x и y, которая не заканчивается дугой $\to y$, то это был бы орпуть $x \leftarrow \cdots \leftarrow y$; но тогда, ввиду существования орпути $z \to \cdots \to y$ (базовая теорема), получили бы цепь $z \to \cdots \to x$, что противоречит условию. Аналогично, к противоречию приводит допущение, что есть цепь между z и y, которая не заканчивается дугой $\to y$. Пункт в) следует из условия $z \in S_{\text{lom}}(x, y)$.

Заметим, что на каждом пути, удовлетворяющем п. в), количество коллайдеров ограничено сверху только кардинальностью $\mathbf{S}$.

Из утверждения 6 следует, что когда между вершиной z, $z \in S_{\text{lom}}(x, y)$, и вершиной x нет ни одной цепи, между ними обязательно есть путь π с одним коллайдером. Существование такого пути вытекает из факта существования цепи $x - \cdots \to y$, которая не проходит через z, и факта существования орпути $z \to \cdots \to y$. Причем единственной коллайдерной вершиной на пути π является либо вершина y, либо какой-то ее предок.

Факт 13. Если верно $z \in S_{\text{lom}}(x, y)$ и $\operatorname{Ds}(x \perp\!\!\!\perp z)$, то $\neg \operatorname{Ds}(x \perp y \perp z)$.

Факт 13 следует из пп. а) и б) утверждения 6.

Утверждение 7 (правило двойного 1-отсечения; double 1-cutting). Если для заданной пары вершин (x, y) существует такая вершина z, что выполняется $\operatorname{Ds}(w \perp z \perp x)$ и $\operatorname{Ds}(w \perp z \perp y)$, то вершина w не входит в состав никакого локально-минимального сепаратора для пары вершин (x, y).

Доказательство. Предположим обратное, т.е. пусть в условиях утверждения будет $w \in S_{\text{lom}}(x, y) = \mathbf{S}$. Если предположить, что вершина w лежит на некоторой цепи между вершинами x и y, то невозможно одновременное выполнение $\operatorname{Ds}(w \perp z \perp x)$ и $\operatorname{Ds}(w \perp z \perp y)$ ни для какой z. Следовательно, w не лежит ни на какой цепи между вершинами x и y. Пусть для определенности имеем случай $\operatorname{Ds}(w \perp\!\!\!\perp x)$. Тогда, согласно базовой теореме, есть орпуть $w \to \cdots \to y$, который не проходит через x. Рассмотрим путь π, между x и y, который закрывается с помощью вершины w. Ясно, что все коллайдеры на участке пути π между вершинами x и w открыты при кондиционировании множества вершин $\mathbf{S} \setminus \{w\}$. Рассмотрим ближайший к вершине w коллайдер $\to q_1 \leftarrow$ на этом участке пути π. Так как он открыт, есть орпуть из q_1 в одну из вершин t_1, входящих в состав $\mathbf{S} \setminus \{w\}$. В свою очередь (согласно базовой теореме), имеется орпуть из t_1 в вершину y. Значит, для выполнения $\operatorname{Ds}(w \perp z \perp y)$ необходимо, чтобы этот орпуть из q_1 в вер-

шину y блокировался с помощью z. (Иначе образуется цепь между вершинами w и y, не блокируемая с помощью z.) Но тогда кондиционирование вершины z открывает коллайдер $\rightarrow q_1 \leftarrow$ на пути π. Рассмотрим следующий коллайдер $\rightarrow q_2 \leftarrow$ на этом участке пути π. Поскольку он тоже открыт при кондиционировании множества вершин $\mathbf{S} \setminus \{w\}$, аналогичные рассуждения приведут к выводу, что есть орпуть из вершины q_2 в вершину z. Следовательно, коллайдер $\rightarrow q_2 \leftarrow$ на пути π также открыт при кондиционировании вершины z. Таким образом, при кондиционировании вершины z открыты коллайдеры $\rightarrow q_1 \leftarrow$ и $\rightarrow q_2 \leftarrow$, следовательно, открыт отрезок пути π между вершинами w и третьим коллайдером $\rightarrow q_3 \leftarrow$. Повторяя рассуждения по этой схеме не более чем $(|\mathbf{S}| - 1)$ раз, приходим к выводу, что из каждого коллайдера q_i на пути π имеется орпуть в вершину z. Следовательно, неверно $\mathrm{Ds}(w \perp z \perp x)$. (В другом случае получим, что неверно $\mathrm{Ds}(w \perp z \perp y)$).

В [12] показано, что утверждение 7 нельзя расширить до правила $[\exists\, z : \mathrm{Ds}(w \perp z \perp x)] \Rightarrow w \notin S_{\mathrm{lom}}(x, y)$. Применение такого правила может привести не только к потере минимального сепаратора, но к потере всех сепараторов для пары вершин (x, y). Утверждение 7 также нельзя расширить до правила вида

$$[\exists\, q, z : \mathrm{Ds}(w \perp q \perp x)\ \&\ \mathrm{Ds}(w \perp z \perp y)] \Rightarrow w \notin S_{\mathrm{lom}}(x, y). \qquad (6)$$

Применение правила (6) может привести только к потере минимального сепаратора, но не к потере всех сепараторов для пары вершин (x, y). Это пояснялось в комментарии к рис. 2. (Если модель содержит скрытые переменные, правило (6) может привести к потере всех сепараторов для пары (x, y).) Утверждение 7 и правило (6) встроены в алгоритм PC.

Для интенсивного сокращения числа кандидатов в сепараторы можно найти и другие возможности. Интуитивно кажется убедительным такое суждение: если одна из вершин пары (x, y) сепарирует другую вершину пары (x, y) от вершины z, то вершина z не является членом никакого $S_{\mathrm{lom}}(x, y)$. Иначе говоря, когда, например, вершина x сепарирует вершину z от вершины y, то можно сказать, что вершина z «отстраняется» от пары (x, y). Формально получаем следующее утверждение.

Утверждение 8. Если $\mathrm{Ds}(z \perp x \perp y)$, то $z \notin S_{\mathrm{lom}}(x, y)$. (При этом можно не уточнять, существует сепаратор для пары вершин (x, y) или нет.)

Доказательство. Допустим противное. Пусть вершина z входит в состав некоторого локально-минимального сепаратора для пары вершин (x, y). Если вершина z лежит на некоторой цепи между вершинами x и y, то понятно, что ни $\mathrm{Ds}(z \perp x \perp y)$, ни $\mathrm{Ds}(z \perp y \perp x)$ не выполняется, т.е. имеем противоречие условию. Остается предположить, что вершина z не лежит ни на какой цепи между вершинами x и y. Тогда z лежит на некотором коллайдерном пути между x и y, так что имеем $\mathrm{Ds}(z \perp \perp x)$ или $\mathrm{Ds}(z \perp \perp y)$. Кроме того, согласно базовой теореме заключаем: а) есть орпуть из z в вершину x, не проходящий через y; или б) есть орпуть из z в вершину y, не проходящий через x. Ввиду условия $\mathrm{Ds}(z \perp x \perp y)$ случай б) отпадает. Значит, есть орпуть из z в вершину x, не проходящий через y; и верно $\mathrm{Ds}(z \perp \perp y)$. Тогда согласно факту 13 из $z \in S_{\mathrm{lom}}(x, y)$ и $\mathrm{Ds}(z \perp \perp y)$ следует $\neg\, \mathrm{Ds}(z \perp x \perp y)$, т.е. получаем противоречие условию.

В [12] утверждение 8 доказано иначе. Можно усилить это утверждение (сделать его более надежным эмпирически) следуюшим образом.

Факт 14 (отстранение вершины; placing aside). Справедливо

$$\mathrm{Ds}(z \perp x \perp y)\ \&\ \neg\, \mathrm{Ds}(z \perp \perp y) \Rightarrow z \notin S_{\mathrm{lom}}(x, y). \qquad (7)$$

В [12] предложено правило «быстрой» идентификации ребра, использующее единственность неотстраненного кандидата.

Согласно утверждению 2 справедлив следующий факт.

Факт 15 (неотстраняемый потенциальный стержень сепаратора). Если существует сепаратор для пары вершин (x, y), то существует такая вершина z $(z \neq x, \quad z \neq y)$, что $\neg \mathrm{Ds}(z \perp\!\!\!\perp x)$, $\neg \mathrm{Ds}(z \perp\!\!\!\perp y)$, $\neg \mathrm{Ds}(z \perp x \perp y)$ и $\neg \mathrm{Ds}(z \perp y \perp x)$.

Определение 7. Вершина z называется потенциальным стержнем сепаратора (potential pivot of separator) для пары вершин (x, y), если $\neg \mathrm{Ds}(x \perp\!\!\!\perp y) \,\&\, \neg \mathrm{Ds}(z \perp\!\!\!\perp x) \,\&\, \neg \mathrm{Ds}(z \perp\!\!\!\perp y) \,\&\, \neg \mathrm{Ds}(z \perp x \perp y) \,\&$ $\& \neg \mathrm{Ds}(z \perp y \perp x)$.

Пусть $\mathbf{U}$ — множество всех вершин графа. Тогда согласно факту 15 получаем следующее правило «быстрой» идентификации ребра (идентификация перемычки; identification of bottleneck):

$$\neg \mathrm{Ds}(x \perp\!\!\!\perp y) \,\&\, (\forall\, z \in \mathbf{U} \setminus \{x, y\} : [\mathrm{Ds}(z \perp\!\!\!\perp x) \vee \mathrm{Ds}(z \perp\!\!\!\perp y) \vee \mathrm{Ds}(z \perp x \perp y) \vee \mathrm{Ds}(z \perp y \perp x)]) \Rightarrow x \text{ — } y, \tag{8}$$

где $\vee$ означает дизъюнкцию.

Как показано в [12], единственный потенциальный стержень сепаратора для (x, y) является обязательным членом любого сепаратора для (x, y) (разумеется, когда нет ребра x — y и когда пустое множество не является сепаратором для (x, y)).

Совокупность предложенных утверждений и фактов позволяет формировать много других полезных правил вывода.

ОБСУЖДЕНИЕ И ОЦЕНКА

Обоснованные выше положения и правила используют в качестве признаков факты зависимости, независимости и некоторые фрагменты структуры модели, а в качестве вывода характеризуют другие фрагменты структуры модели. Смысл (назначение) большинства предложенных положений и фактов следующий:

- распознавание ребер;
- распознавание существования сепаратора (т.е. отсутствия ребра);
- идентификация невхождения вершины в состав локально-минимального сепаратора.

Для этого установлены необходимые требования к членам (локально-минимального) сепаратора. Достаточные требования к членам локально-минимального сепаратора (разумеется, когда соответствующий сепаратор еще не найден и не верифицирован) предполагают знание, что соответствующее ребро отсутствует. Иногда такое знание может быть выведено индуктивно. Тогда можно индуктивно идентифицировать вершину как члена локально-минимального сепаратора. Например, достаточные условия для $z \in S_{\mathrm{lom}}(x, y)$ можно сформировать как такое сочетание: 1) правило непоглощения; 2) правило единственного общего близкого или правило множества изолированных общих близких.

Предложенные средства (в частности, необходимые требования к членам локально-минимального сепаратора) следует встроить в алгоритм вывода структуры с помощью нескольких принципов. Сохраняем базовый принцип вывода, принятый в алгоритме PC: если испытаны и отвергнуты все подмножества из множества правдоподобных членов сепаратора для пары (x, y), то существует ребро x — y. Однако эффективность работы этого принципа повышается благо-

даря введению дополнительных требований к кандидатам в члены сепаратора и к составу сепаратора. В частности, в состав каждого сепаратора должен входить хотя бы один потенциальный стержень сепаратора. Эти требования и правила часто позволяют отсеять значительно больше кандидатов в сепаратор, а иногда — даже всех кандидатов. Тогда поиск сепаратора резко упрощается, а иногда прекращается на раннем этапе. В пределе получаем новые принципы вывода.

Резолюция смежности. Если не осталось ни одного неотсеянного кандидата в сепараторы для пары ассоциированных вершин (x, y), то существует ребро x — y.

Этот принцип покрывается следующим.

Экспресс-резолюция смежности. Если для пары вершин (x, y) не осталось ни одного забракованного потенциального стержня сепаратора, то существует ребро x — y.

Дополнительно показано, как может работать концептуально новый принцип вывода, который не требует ни поиска соответствующего сепаратора, ни анализа и отсеивания кандидатов в сепаратор.

Резолюция несмежности. Если набор фактов свидетельствует о невозможности ни дуги (орпути) $x \to y$, ни дуги (орпути) $y \to x$, то ребро x — y отсутствует.

Совокупность предложенных средств открывает перспективы значительно повысить вычислительную эффективность идентификации структуры модели.

Напомним, что все полученные результаты предполагают отсутствие строго ориентированных циклов. Скрытые переменные (отображаемые дву-ориентированными ребрами) с определенными ограничениями допустимы для некоторых положений, однако недопустимы для утверждений 1, 3, 4 и фактов 6, 8–10, 12.

Потенциальные возможности некоторых предложенных инструментов можно продемонстрировать на примере модели, показанной на рис. 4. Этот граф состоит из 17 вершин и 21 ребра. Каждая из вершин x и y имеет 13 безусловно зависимых вершин (не считая самих x и y), причем две из них — совместные смежные для x и y. Если применить к этому примеру тактику алгоритма PC, то для идентификации ребра x — y будет выполнен сложный перебор с использованием тестов независимости высокого порядка. С помощью предложенных средств ребро x — y идентифицируется на основе фактов (тестов) условной независимости только нулевого и первого порядка. В этом примере достаточно применить правило «чужого гена», правило двойного 1-отсечения и правило «отстранения».

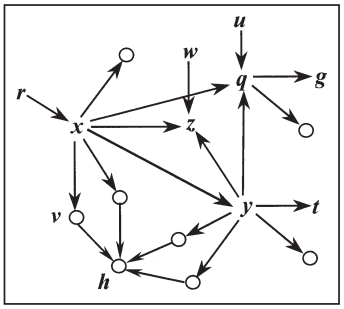

Рис. 4

Предложенные положения и правила в том виде, как они сформулированы выше, непосредственно пригодны для вывода из совокупности (базы) знаний о системе зависимостей. Другими словами, эти средства служат для компиляции (синтеза) модели из отдельных знаний. (Заметим, что знания в форме каузального порядка переменных сделают некоторые предложенные инструменты излишними.) Кроме того, данные правила применимы для анализа модели, в частности для планирования схемы рассуждений на заданной модели, т.е. для вывода от свидетельств к целевым переменным (в экспертных системах). В таких рассуждениях информация распространяется по структуре модели через сепараторы. А именно, если y — целевая переменная, а переменная x входит в состав свидетельства, то (при точном выводе) информация будет проходить через каждую

вершину $z \in S_{\text{lom}}(x, y)$ каждого локально-минимального сепаратора $S_{\text{lom}}(x, y)$.

Поскольку предложенный инструментарий в первую очередь предназначен для методов вывода структуры модели из статистических данных, его нужно настроить для этих целей. Графовые предикаты следует заменить эмпирическими «двойниками» (counterparts) согласно эквивалентности (3). Однако при практическом использовании эмпирических версий введенных утверждений и правил возникнут трудности ввиду проблематичности полной версии предположения необманчивости (2). Предложенные утверждения и правила более тонкие, чем большинство известных. Они характеризуют одни сепараторы на основе знаний других сепараторов и зависимостей и опираются на более строгие версии предположения необманчивости, чем (3) и (4). В частности, эмпирические «двойники» утверждений 1, 2a, 7 и фактов 5, 6, 10–13, 15 опираются на свойства путей (а не только ребер) и требуют выполнения предположения необманчивости в значительно большем объеме, чем (4). Поэтому прямолинейное применение эмпирических «двойников» правил при выводе из небольшой выборки данных может привести к потере сепараторов и идентификации ложных ребер. Однако риск ограничивается благодаря тому, что используются (как признаки) отношения (не)зависимости только нулевого и первого порядка.

Предположение необманчивости может нарушаться даже в случае очень большой (асимптотической) выборки данных. Известны примеры моделей, в которых даже базовый принцип (4) дает ошибки. На практике часто выборка данных — небольшая или малая. Тогда выборочное распределение может значительно отклоняться от генеративного. В результате расширяется сфера нарушения предположения необманчивости. Увеличение кардинальности условий тестов независимости еще больше обостряет проблему ненадежности.

Эмпирический факт зависимости — более удобное свидетельство для выводов, чем эмпирическая независимость. В частности, $\neg \Pr(x \perp \perp y)$ есть довольно робастное свидетельство существования цепи. Импликация $\neg \Pr(x \perp \mathbf{S} \perp y) \Rightarrow$ $\Rightarrow \neg \operatorname{Ds}(x \perp \mathbf{S} \perp y)$ верна, а импликация $\Pr(x \perp \mathbf{S} \perp y) \Rightarrow \operatorname{Ds}(x \perp \mathbf{S} \perp y)$ — нет. Эмпирическая независимость не гарантирует отсутствия цепи, но свидетельствует о слабости (возможной) зависимости. Довольно надежна редуцированная версия: $\Pr(x \perp \mathbf{S} \perp y) \Rightarrow \neg (x — y)$.

Надежность вывода на основе эмпирических свидетельств можно повысить за счет контрастных тестов зависимости-независимости:

- $\Pr(z \perp x \perp y) \,\&\, \neg \Pr(z \perp \perp y)$ — актуальная 1-сепарация [15].
- $\Pr(z \perp \perp y) \,\&\, \neg \Pr(z \perp x \perp y)$ — провокация зависимости [13].

Благодаря актуальной сепарации можно избежать ошибочного отстранения вершины в ситуациях типа «слабая провоцированная зависимость», что можно проиллюстрировать следующим примером. Пусть имеем гауссову сеть, описываемую уравнениями:

$$x = \varepsilon_1; \quad w = \varepsilon_2; \quad z = x + 0{,}8w + \varepsilon_3;$$

$$y = 0{,}7z + aw + \varepsilon_4.$$

Пусть все члены ошибок ε_i взаимно независимы и имеют стандартное отклонение $\sigma = 2$. В этой модели единственным сепаратором для (x, y) является $\{z, w\}$. Согласно факту 13 будет $\neg \operatorname{Ds}(x \perp y \perp w)$, поэтому w не должна отстраняться. Но в небольших выборках ошибка отстранения возможна. Предусловием такой ошибки является определенный набор соотношений для выборочных значений коэффициентов корреляции, а именно, должно быть: $|\rho_{xw|y}| \leq |\rho_{wy}|$. Кроме того, должно быть: $|\rho_{xw|y}| \leq |\rho_{xy}|$; $|\rho_{xw|y}| \leq |\rho_{xy|z}|$; $|\rho_{xw|y}| \leq |\rho_{wz}|$;

$|\rho_{xw|y}| \le |\rho_{xz}|$; $|\rho_{xw|y}| \le |\rho_{zy}|$. При указанных в уравнениях значениях параметров эти неравенства выполняются для $-1{,}35 < a < -0{,}29$. Предположим, структуральные коэффициенты имеют именно такие выборочные значения, как записано выше, и при этом $a = -0{,}9$. Тогда получим следующие выборочные оценки значений коэффициентов (частной) корреляции: $\rho_{xw|y} = 0{,}133$; $\rho_{wy} = -0{,}235$; $\rho_{xy} = 0{,}484$; $\rho_{xz} = 0{,}615$; $\rho_{zy} = 0{,}480$; $\rho_{wz} = 0{,}492$; $\rho_{xy|z} = = 0{,}272$. Легко представить, что, при определенном размере выборки, в результате тестирования будет принята независимость $\Pr(x \perp y \perp w)$ (синдром слабой провоцированной зависимости), в то время как для остальных перечисленных отношений — нет. Тогда переменная w будет исключена из числа кандидатов в сепаратор и потеряется единственный сепаратор. (Разумеется, это упрощенное объяснение. Для корректности нужно сравнивать значения z-статистики Фишера.)

Версия (7) правила защитит от ошибки в описанной ситуации. Это объясняет целесообразность усиления правила «отстранения» инструментом актуальной сепарации.

Надежность эмпирической версии «отстранения» (7) можно аргументировать следующим образом [12]. Факты $\mathrm{Ds}(z \perp x \perp y) \,\&\, \neg\, \mathrm{Ds}(z \perp\!\!\!\perp y)$ свидетельствуют, что после закрытия цепей через вершину x уже не остается сильных цепей между вершинами y и z. Тогда можно заключить, что все возможные цепи между вершинами y и x, идущие через z, будут еще слабее, поскольку включают в себя указанные слабые цепи как часть. Однако эта аргументация не учитывает взаимодействия путей, так что при малых выборках данных и неудачных сочетаниях параметров риск ошибки остается.

Ненадежной будет эмпирическая версия правила «чужого гена» (см. факт 11). Но с помощью провокации зависимости можно получить довольно надежное правило идентификации обоюдоблизкой коллайдерной вершины (non-ancestral common covariate):

$$\exists\, w{:}[\neg \Pr(w \perp\!\!\!\perp z) \,\&\, \Pr(w \perp\!\!\!\perp x) \,\&\, \neg \Pr(w \perp z \perp x) \,\&$$

$$\&\, \Pr(w \perp\!\!\!\perp y) \,\&\, \neg \Pr(w \perp z \perp y)] \Rightarrow z \notin S_{\mathrm{lom}}(x, y). \qquad (9)$$

С помощью (9) для примера на рис. 4 вершины u и w можно идентифицировать как «чужие гены» для вершин x и y и исключить q и z из поиска сепаратора.

Отметим, что предложенные инструменты предоставляют возможности в процессе вывода комбинировать априорные знания (о фрагментах графа) и эмпирические свидетельства [12].

ЗАКЛЮЧЕНИЕ

Установлены необходимые требования к членам сепараторов, а также признаки существования или отсутствия ребер. Данные требования к членам сепараторов позволяют отсеять некоторых кандидатов в сепаратор и тем самым упростить задачу идентификации структуры модели. Предложенные и обоснованные утверждения и правила позволяют (адаптивно) оптимизировать поиск сепараторов в процессе анализа или синтеза АОГ-моделей. Эти средства имеют следующие преимущества:

- ускоряют и упрощают идентификацию существования или отсутствия ребер;
- способствуют раннему сокращению списков кандидатов в сепараторы;
- обеспечивают выявление минимальных сепараторов;

• способствуют уменьшению числа тестов независимости при выводе из данных;

• способствуют снижению порядка (ранга) тестов условной независимости и упрощению запросов к базе данных;

• поддерживают использование априорных знаний о структуре модели при окончательном выводе структуры АОГ-модели.

В работе показаны новые возможности и аналитические средства, которые повышают эффективность индуктивного вывода структур вероятностных моделей зависимостей.

СПИСОК ЛИТЕРАТУРЫ

1. Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference. — San Mateo: Morgan Kaufmann, 1988. — 552 p.
2. Scheines R., Spirtes P., Glymour C. et al. The TETRAD project: Constraint based aids to causal model specification // Multivar. Behavior. Res. — 1998. — **33**, N 1. — P. 65–118.
3. Neapolitan R.E. Learning Bayesian networks. — Englewood Cliffs: Prentice Hall, 2003. — 674 p.
4. Pearl J. Causality: models, reasoning, and inference. — Cambridge: Cambridge Univ. Press, 2000. — 526 p.
5. Scheines R., Spirtes P., Glymour C., Meek C. TETRAD II: Tools for discovery. — Hillsdale, NJ: Lawrence Erlbaum Assoc., 1994.
6. Verma T., Pearl J. Causal networks: semantics and expressiveness / R. Shachter, T.S. Levitt, L.N. Kanal (Eds.) // Uncertainty in Artificial Intelligence. — Amsterdam: Elsevier Sci. Publ., North-Holland, 1990. — **4**. — P. 69–76.
7. Scheines R. An introduction to causal inference // Causality in Crisis? / V. McKim and S. Turner (Eds.). — Notre Dame: Univ. of Notre Dame Press, 1997. — P. 185–200.
8. Андон Ф.И., Балабанов А.С. Структурные статистические модели: инструмент познания и моделирования // Систем. дослід. та інформ. технології. — 2007. — № 1. — С. 79–98.
9. Chickering D.M., Meek C., Heckerman D. Large-sample learning of Bayesian networks is NP-hard // Proc. of 19th Conf. on Uncertainty in Artificial Intelligence. — Acapulco, Mexico: Morgan Kaufmann, 2003. — P. 124–133.
10. Meek C. Strong-completeness and faithfulness in belief networks / S. Hanks, P. Besnard (Eds.) // Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (Montreal, QU). — San Mateo, CA: Morgan Kaufmann, 1995. — P. 411–418.
11. Балабанов А.С. Восстановление структур систем вероятностных зависимостей из данных. Аппарат генотипов переменных // Проблемы управления и информатики. — 2003. — № 2. — С. 91–99. (see: http://www.begellhouse.com/journals/)
12. Балабанов О.С. Правила пцдбору сепараторцв у баєсівських мережах // Проблеми програмування. — 2007. — № 4. — С. 22–33.
13. Балабанов А.С. К выводу структур моделей вероятностных зависимостей из статистических данных // Кибернетика и системный анализ. — 2005. — № 5. — С. 19–31.
14. Tian J., Paz A., Pearl J. Finding minimal *d*-separators: Techn. Rep. / UCLA, Computer Sci. Dep, CA. — R-254. — Los Angeles, 1998. — 15 p.
15. Балабанов А.С. Индуктивный метод восстановления монопотоковых вероятностных графовых моделей зависимостей // Проблемы управления и информатики. — 2003. — № 5. — С. 75–84.

Поступила 26.06.2008