Компьютерная Математика

Экспертные системы, методы индуктивного вывода

УДК 007:681.3.00

А.С. БАЛАБАНОВ

О ЛОГИКЕ НЕЗАВИСИМОСТИ КАУЗАЛЬНЫХ ДИАГРАММ

Анализируются свойства рекурсивных каузальных диаграмм (АОГ-моделей). Показано, что множество парных безусловных независимостей влечет ограничения на допустимые условные независимости и определяет некоторые фрагменты структуры диаграммы.

Введение. Графовые модели вероятностных зависимостей – строгий язык представления знаний о зависимостях в условиях неопределенности, использующий многомерный статистический анализ, аппарат графов, теорию информации и логику. Наиболее популярным классом графовых моделей стали модели, структурированные ациклическими ориентированными графами (АОГ-модели). Достоинства АОГ-моделей – компактное представление систем зависимостей и вычислительная эффективность вероятностного вывода от свидетельств. АОГ-модели привлекают своей способностью отображать причинно-следственные связи, поэтому они также известны как каузальные диаграммы [1–3]. Иногда в каузальных диаграммах допускают латентные переменные. Здесь рассматриваются модели без латентных переменных, т.е. рекурсивные каузальные диаграммы. В таких диаграммах нет двунаправленных дуг. Интерес представляют независимости только для пар переменных. Как только специфицировано множество парных фактов безусловной независимости, для существования непротиворечивой диаграммы необходимо, чтобы выполнялись (или не выполнялись) соответствующие условные независимости. Некоторые импликации отношений независимости можно получить с помощью полуграфоидных аксиом [2–5]. Однако эти аксиомы используют информацию фрагментарно и никак не учитывают специфику каузальных диаграмм – ориентированность связей, отсутствие ориентированных циклов и скрытых переменных. Будем опираться на топологические свойства каузальных диаграмм и системно использовать множество безусловных независимостей двух переменных.

Постановка задачи. Пусть специфицировано все множество безусловных независимостей пар переменных. Требуется построить такую каузальную диаграмму (или предсказать ее необходимые фрагменты), чтобы совершенно отобразить все Марковские свойства (согласно критерию *d*-сепарации). Требуется вывести условные независимости и структурные фрагменты диаграммы, которые должны присутствовать в любой диаграмме, удовлетворяющей заданным условиям. То есть необходимо выявить нетривиальные импликации заданного множества безусловных независимостей и общих конструктивных ограничений каузальных диаграмм.

Базовые определения. Если в графе G есть дуга $x \to y$, то вершина x называется "родителем" вершины y, а вершина y – ребенком x. Ребро – это дуга, ориентация которой неизвестна или игнорируется. Ребро обозначается x—y, а отсутствие ребра – как $\neg(x$—$y)$. Путь в орграфе – это последовательность инцидентных ребер без повторения вершин. Как исключение, первая и последняя вершины пути (из трех или более ребер) могут совпадать, и такой путь называют циклом. Орпуть (униориентированный путь) – это путь, на котором все ребра ориентированы в одном направлении ($x \to \cdots \to y$). Ациклический ориентированный граф (АОГ) – это орграф без униориентированных циклов. Далее под графом подразумеваем АОГ. Ввиду ацикличности графа и униориентированности каждого ребра любой АОГ содержит корни. Корень орграфа – это вершина, не имеющая ни одной входящей дуги. Вершина x называется предком вершины y в орграфе, если есть орпуть из x в y (в частности, путь нулевой длины). Отношение предок-потомок является рефлексивным транзитивным замыканием отношения родитель-ребенок. Обозначим $F(x)$ множество родителей x.

Ввиду взаимно однозначного соответствия термины переменная (модели) и вершина (графа) употребляются как взаимозаменяемые. АОГ-модель определяется как (D, θ), где D – АОГ, а θ – совокупность локально заданных параметров $p(x|F(x))$. АОГ-модель определяет совместное распределение вероятностей переменных согласно правилу Байеса [1–3].

Определение 1. Коллайдером в графе называется фрагмент вида $x \to y \leftarrow z$. Коллайдер $x \to y \leftarrow z$ называется шунтированным (экранированным), если в графе есть ребро x—z. Цепь (бесколлайдерный путь) – это путь, не содержащий ни одного коллайдера.

Ключевую роль в теории АОГ-моделей играет критерий *d*-сепарации. Этот критерий объединяет каузальное Марковское предположение и полуграфоидные аксиомы независимости [1–5].

Определение 2. Путь π в АОГ-модели называют *d*-закрытым с помощью (кондиционирования) множества вершин $\mathbf{S}$, если и только если

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

б) на пути π есть хотя бы один коллайдер $\to y \leftarrow$, причем $y \notin \mathbf{S}$ и не существует никакой вершины $z \in \mathbf{S}$, такой, что есть орпуть из y в z.

Множество вершин $\mathbf{Z}$ d-сепарирует вершины x и y, если и только если все пути между вершинами x и y являются d-закрытыми с помощью множества вершин $\mathbf{Z}$. Будем обозначать такую d-сепарацию предикатом $\mathrm{Ds}(x \perp \mathbf{Z} \perp y)$. Когда $\mathbf{Z}=\{\}$, будем писать такой предикат как $\mathrm{Ds}(x \perp\!\!\perp y)$. Если хотя бы один путь между x и y не является d-закрытым с помощью $\mathbf{Z}$, то говорят, что вершины x и y d-соединены (d-зависимы), т. е. $\neg\mathrm{Ds}(x \perp \mathbf{Z} \perp y)$.

Если в АОГ предикат $\mathrm{Ds}(x \perp \mathbf{Z} \perp y)$ истинен, то множество вершин $\mathbf{Z}$ называют сепаратором, или d-сепаратором, для пары (x, y). Первичное свойство сепарации в АОГ связывает идентификацию ребер и независимостей:

$$(x—y) \ \Leftrightarrow \forall \mathbf{Z}(x, y \notin \mathbf{Z}) : \neg\mathrm{Ds}(x \perp \mathbf{Z} \perp y). \tag{1}$$

Условную независимость переменной x от переменной y при кондиционировании множества переменных $\mathbf{Z}$ (x,$y$$\notin$$\mathbf{Z}$) будем обозначать $\Pr(x \perp \mathbf{Z} \perp y)$. Для дискретных переменных условная независимость означает $p(xy|\mathbf{Z}) = p(x|\mathbf{Z}) \cdot p(y|\mathbf{Z})$ для всех значений. Безусловную независимость будем записывать в форме $\Pr(x \perp\!\!\perp y)$.

Известно [1–3, 5], что критерий d-сепарации определяет Марковские свойства АОГ-модели (т. е. условные независимости, выполняющиеся при любой параметризации модели):

$$\forall \mathbf{Z}, x, y \notin \mathbf{Z} : \mathrm{Ds}(x \perp \mathbf{Z} \perp y) \Rightarrow \Pr(x \perp \mathbf{Z} \perp y). \tag{2}$$

Предположение каузальной необманчивости распределения вероятностей переменных модели относительно АОГ выражается как обратная импликация по сравнению с (2). Объединяя это предположение с (2), получаем

$$\forall \mathbf{Z}(x, y \notin \mathbf{Z}) : \mathrm{Ds}(x \perp \mathbf{Z} \perp y) \Leftrightarrow \Pr(x \perp \mathbf{Z} \perp y). \tag{3}$$

При этом говорят, что АОГ-модель совершенно отображает парные Марковские свойства. В данной работе рассматриваются именно совершенные АОГ-модели. То есть принимается, что множество фактов (условной) независимости точно и полно определяется критерием d-сепарации. В частности, две вершины безусловно независимы тогда и только тогда, когда между ними существует цепь.

Пусть задано множество безусловных независимостей, выполняющихся в модели. Тогда можно установить некоторые рамки для множества более сложных (условных) независимостей, которым должна удовлетворять АОГ-модель, чтобы совершенно отображать Марковские свойства.

Аппарат генотипов переменных. Поскольку множество безусловно зависимых переменных может быть большим, предлагается использовать более экономное (компактное) их представление через понятие генотипа переменной. Каждая корневая переменная является предком для всех своих потомков, поэтому естественно назвать ее «геном» (или приписать ей ген).

Определение 3. *Генотип* $G(x)$ переменной (вершины) x – это множество всех корневых предков переменной x. (Каждая переменная является предком, в частности, и самой себя.) Популяция переменных ─ это множество переменных, которые имеют одинаковый генотип. Две переменные x и y называются родственными, если они имеют хотя бы один общий ген, т.е. $G(x) \cap G(y) \neq \{\}$. Легко убедиться, что справедлива эквивалентность

$$G(x) \cap G(y) = \{\} \Leftrightarrow \mathrm{Ds}(x \perp\!\!\perp y) \Leftrightarrow \mathrm{Pr}(x \perp\!\!\perp y). \qquad (4)$$

Таким образом, генотипы точно отображают все безусловные независимости.

Набором безусловных независимостей переменной x назовем множество $I(x) = \{y | \mathrm{Ds}(x \perp\!\!\perp y)\}$. Генотипы переменных легко вывести из отношения безусловной независимости. Нижеследующая процедура является уточнением процедуры “Геном-1” из [6]. На входе полностью специфицировано множество фактов безусловной независимости вершин (переменных) модели.

1. Для каждой переменной x сформировать ее набор безусловных независимостей $I(x)$. При этом x назовем порождающей переменной для $I(x)$.

2. Упорядочить множество наборов $I(x)$ по кардинальности.

3. Выявить все совпадающие наборы. Взять по одному набору (представителю) из каждого подмножества равных наборов. Составить список L из наборов-представителей и упорядочить L по кардинальности.

4. В списке наборов L найти все отношения поглощения (включения).

5. Выбрать из списка L все те наборы, которые не поглощаются другими, и образовать из них список L^*. Для каждого набора из списка L^* приписать порождающей переменной уникальное имя гена. (Таким именем можно объявить имя самой порождающей переменной.) Получаем набор генов R.

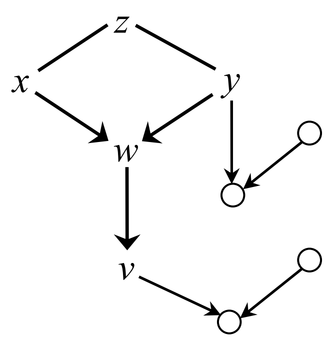
6. Порождающим переменным всех наборов из списка L, которые совпадают с наборами из L^*, приписать те же имена генов. Таким образом, получаем все одногенные популяции переменных.

7. Определить генотипы остальных переменных. Множество отсутствующих генов определяется как $AG(x) = \{y \in R | \mathrm{Ds}(x \perp\!\!\perp y)\}$. Тогда

$$G(x) = R \setminus AG(x).$$

Согласно определению, в роли генов выступают корневые вершины графа. Однако не всегда можно точно определить корневую вершину. Известно, что если одногенная популяция имеет структуру дерева, то корень в ее составе не распознается фактами независимости. Но когда в составе одногенной популяции есть не шунтированный коллайдер, выясняется, что ни коллайдерная переменная, ни ее потомок не могут быть корнем. Ввиду этого можно уточнить корни (и гены), выданные процедурой “Геном-1”. Например, пусть модель (рисунок), содержит одногенную популяцию $\{x, y, z, w, v\}$. Процедура “Геном-1” могла бы выдать в качестве гена переменную w или v. Но ввиду того, что $\mathrm{Ds}(x \perp z \perp y)$ и других фактов выясняется, что ни w, ни v не могут быть корнями. Тогда можно скорректировать имена и объявить геном переменную x или y, или z. (Для дальнейших выкладок эта коррекция не существенна. Для всех вершин за пределами данной одногенной популяции не имеет значения, какая из вершин популяции – корень.)

Легко видеть, что в АОГ с k корнями выполняется не менее $\sum_{i=1}^{k} r_i \left(\sum_{j=i+1}^{k} r_j \right)$ парных безусловных независимостей, где r_i – кардинальность i-й одногенной популяции. В АОГ с двумя корнями будет точно $r_1 r_2$ безусловных независимостей. В частном случае, когда АОГ имеет только один корень, нет ни одной безусловной независимости.



РИСУНОК

Генотипом множества переменных (вершин) назовем объединение генотипов всех элементов множества.

Импликации отношения родственности. В аргументации неявно применяется операция вида «бесколлайдерная стыковка». Если существует цепь между вершинами x и y и орпуть вида $y \to \cdots \to z$, то существует цепь между вершинами x и z. Это следует прямо из критерия d-сепарации. (Если при таком соединении образуется цикл на промежуточной вершине, цикл следует выбросить из итоговой цепи.) Установим простые, но полезные факты для генотипов.

Факт 1. По дуге (в направлении стрелки) наследуются все гены. Если есть дуга $x \to y$, то $G(x) \subseteq G(y)$. Следовательно, по униориентированному пути наследуются все гены, т. е. если есть $x \to \cdots \to y$, то $G(x) \subseteq G(y)$, и для всех промежуточных вершин z на этом пути $G(x) \subseteq G(z) \subseteq G(y)$.

На любой цепи есть старшая вершина x, т. е. такая вершина, что на этой цепи нет ни дуги $\to x$, ни дуги $x \leftarrow$. В случае униориентированного пути старшей является начальная вершина. В других случаях цепи старшей является промежуточная вершина, имеющая на этой цепи дуги $\leftarrow x \to$.

Факт 2. Если на цепи старшей является вершина x, то для всех других вершин y на этой цепи верно $G(x) \subseteq G(y)$.

Следствия из фактов 1 и 2:

если вершина z лежит на некоторой цепи между x и y, то $G(z) \subseteq G(x)$ или $G(z) \subseteq G(y)$;

если $\neg G(x) \subseteq G(y)$, то невозможны ни дуга $x \to y$, ни орпуть $x \to \cdots \to y$;

если $\neg G(x) \subseteq G(y)$ и $\neg G(x) \supseteq G(y)$, то невозможно ребро x—y и невозможен никакой орпуть между x и y.

Факт 3. Если в модели есть только одна двухгенная переменная y с генотипом $G(y)=\{x,z\}$, а одногенные популяции с генотипами $\{x\}$ и $\{z\}$ соответственно состоят из единственных переменных x и z, то существует коллайдер $x \to y \leftarrow z$.

Действительно, в этих условиях не существует никаких промежуточных переменных, через которые мог бы пройти орпуть из x в y и орпуть из z в y.

Обобщая эту аргументацию, получаем.

Факт 4. Пусть задана многогенная переменная y, причем $a \in G(y)$. Если в модели среди всех переменных z с числом генов $2 \le |G(z)| \le |G(y)|$ нет ни одной такой, что $G(z) \subseteq G(y)$ и $a \in G(z)$, то существует дуга вида $x \to y$, выходящая из одногенной популяции с генотипом $G(x)=\{a\}$.

Среди множества фактов условной независимости аналитику прежде всего интересны более простые, «компактные» отношения. Это формализовано понятием локально-минимального сепаратора [7]. Известно [2, 7], что из каждой вершины каждого локально-минимального сепаратора для пары (x, y) существует орпуть в вершину x или в вершину y. Следовательно, согласно факту 1, справедливы такие правила для независимостей.

Правило 1. Если существует некоторое множество $\mathbf{Z}$, такое, что $\mathrm{Ds}(x \perp \mathbf{Z} \perp y)$, т. е. отсутствует ребро $x—y$, то также существует некоторое множество $\mathbf{Z^*}$, такое, что $\mathrm{Ds}(x \perp \mathbf{Z^*} \perp y)$, причем $G(\mathbf{Z^*}) \subseteq G(x)$ или $G(\mathbf{Z^*}) \subseteq G(y)$.

Следствие из правила 1. Если отсутствует ребро $x—y$ и для вершины w выполняется $\neg G(w) \subseteq G(x)$ и $\neg G(w) \subseteq G(y)$, то существует некоторое множество $\mathbf{Z}$, такое, что $\mathrm{Ds}(x \perp \mathbf{Z} \perp y)$, причем $w \notin \mathbf{Z}$.

Правило 2. Если нет ребра $x—y$, то существует некоторое множество $\mathbf{Z}$, такое, что $\mathrm{Ds}(x \perp \mathbf{Z} \perp y)$, причем для любого такого $\mathbf{Z}$: $G(\mathbf{Z}) \supseteq G(x) \cap G(y)$.

Покажем, что иногда можно вывести существование сепаратора с более жестким ограничением на генотип.

Правило 3. Если $\neg G(x) \subseteq G(y)$ и $\neg G(x) \supseteq G(y)$, то существует некоторое множество $\mathbf{Z}$, такое, что $\mathrm{Ds}(x \perp \mathbf{Z} \perp y)$, причем $G(\mathbf{Z}) = G(x) \cap G(y)$.

Из условия следует, что невозможно ребро $x—y$ и невозможен никакой орпуть между x и y. Очевидно, что $G(x) \cap G(y)$ – наименьший генотип любого сепаратора для пары вершин (x,y). Существование такого сепаратора можно показать так. Старшая вершина на каждой цепи между x и y является промежуточной. Объединение всех этих старших вершин даст искомый сепаратор $\mathbf{Z}$.

Заключение. Предложены новые правила вывода, позволяющие получать некоторые условные независимости и некоторые обязательные фрагменты или характеристики каузальной диаграммы, имплицируемые заданным множеством парных безусловных независимостей. Предложенный аппарат генотипов и правила помогают на основе спецификации безусловных независимостей синтезировать структуру модели, совершенно представляющую Марковские свойства.

О.С. Балабанов

ПРО ЛОГІКУ НЕЗАЛЕЖНОСТІ КАУЗАЛЬНИХ ДІАГРАМ

Аналізуються властивості рекурсивних каузальних діаграм (АОГ-моделей). Показано, що множина парних безумовних незалежностей тягне обмеження на припустимі умовні незалежності та визначає деякі фрагменти структури діаграми.

O.S. Balabanov

ON INDEPENDENCY LOGIC OF CAUSAL DIAGRAMS

Some properties of recursive causal diagrams (DAG models) are analyzed. We show that fully specified set of pairwise unconditional independencies entails certain restrictions on set of conditional independencies in the model and determines certain fragments of the model.

1. *Pearl J.* Causality: Models, Reasoning, and Inference. – Cambridge Univ. Press, 2000. – 526 p.
2. *Lauritzen S.L.* Causal inference from graphical models / O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg (eds.). // Complex Stochastic Systems / London: Chapman and Hall, 2000. – P. 63–107.
3. *Verma T.* Causal networks: semantics and expressiveness / Technical Report R-65-I. – Cognitive Systems Laboratory, UCLA, Los Angeles, CA. – June 1987. – 10 p.
4. *Dawid A. P.* Conditional independence / S. Kotz, C. B. Read and D. L. Banks. (eds.) // Encyclopedia of Statistical Sciences. – 1998. – **2**. – P.146–155.
5. *Pearl J., T. Verma*. The Logic of Representing Dependencies by Directed Graphs // Proc. of 6th National Conf. on Artificial Intelligence (AAAI-87), July 13–17, 1987. – Seattle, WA: AAAI Press, 1987. – P. 374–379.
6. *Балабанов А.С.* Восстановление структур систем вероятностных зависимостей из данных. Аппарат генотипов переменных // Проблемы управления и информатики. – 2003. – № 2. – С. 91 – 99.
7. *Балабанов А.С.* Минимальные сепараторы в структурах зависимостей. Свойства и идентификация // Кибернетика и системный анализ. – 2008. – № 6. – С. 17–32.

Получено 09.05.2009

Об авторе:

Балабанов Александр Степанович,
кандидат технических наук, старший научный сотрудник
Института программных систем НАН Украины.