

УДК 004.8

МАТРОИДНОЕ ПРЕДСТАВЛЕНИЕ СЕМЕЙСТВА
ГРАФОВ СМЕЖНОСТИ НАД НАБОРОМ ФРАГМЕНТОВ ЗНАНИЙ

В.В. Опарин, А.А. Фильченков, А.В. Сироткин, А.Л. Тулупьев

В работе исследуется одна из задач, возникающих при автоматическом обучении баз фрагментов знаний с неопределенностью, представленных в виде алгебраических байесовских сетей – построение графа смежности как глобальной (вторичной) структуры сети по ее первичной структуре. Установлено, что семейство графов смежности при заданной первичной структуре сети характеризуется матроидом специального вида, что влечет ряд других заключений: множества минимальных и нередуцируемых графов смежности совпадают, минимальный граф смежности допускает построение с помощью жадного алгоритма, число ребер минимального графа смежности выражается через число ребер максимального графа смежности и ранг матроида.

Ключевые слова: базы фрагментов знаний, алгебраические байесовские сети, автоматическое обучение, глобальная структура, матроид, граф смежности.

Если ваша жизнь — матроид, вы можете пользоваться жадным алгоритмом: он будет работать. Иначе — не стоит.

Проф. А.А. Шальто

Введение

Одной из актуальных и активно исследуемых проблем в области проектирования, разработки и эксплуатации систем, основанных на знаниях, является их автоматическое обучение (машинное обучение, machine learning). Для каждого класса систем, основанных на знаниях, формируется своя совокупность задач автоматического обучения; в случае алгебраических байесовских сетей (АБС) [1–4], относящихся к классу логико-вероятностных графических моделей баз фрагментов знаний с неопределенностью, эта совокупность делится на несколько составляющих, среди которых выделяется задача обучения глобальной структуры АБС (или вторичной структуры АБС) по известной первичной структуре этой сети [5, 6].

Первичная структура АБС – это просто набор фрагментов знаний, где каждый фрагмент знаний является идеалом конъюнктов со скалярными или интервальными оценками вероятности истинности. Вторичная структура АБС (глобальная структура АБС) – это совокупность связей между фрагментами знаний, представленная в виде графа смежности (ГС) [1–4]. Одной и той же первичной структуре может соответствовать несколько графов смежности [1, 4, 7–9]; для реализации ряда алгоритмов логико-вероятностного вывода (проверка и поддержание непротиворечивости АБС, апостериорного вывода) выбор графа смежности существен, причем наиболее удачным выбором графа смежности для формирования вторичной структуры АБС является ациклический граф смежности (дерево смежности) [2]. Отсюда возникает потребность как в исследовании семейства графов смежности, сформированных над одним и тем же набором фрагментов знаний, так и в выделении из этого семейства нередуцируемых и минимальных (по числу ребер) графов смежности (НГС и МГС соответственно), поскольку именно они могут оказаться ациклическими. Таким образом, описание семейства графов смежности, а также нередуцируемых и минимальных графов смежности становится целью настоящей работы.

Постановка задачи и определения

Пусть задан конечный алфавит $A = \{x_i\}_{i=1}^m$ и множество вершин $V = \{v_i\}_{i=1}^n$. На множестве вершин задана весовая функция $W: V \rightarrow 2^A$. Значение функции на конкретной вершине v будем называть *весом вершины* и обозначать его через W_v . В общем случае любое подмножество алфавита A будем называть *весом*.

Пусть на множестве вершин V построен некоторый граф $G = \langle V, E \rangle$. Рассмотрим вершины $u, v \in V$. Пусть $W_u \cap W_v = q \neq \emptyset$.

Определение 1. Назовем вершины u, v *магистрально связными*, если существует путь P из u в v , веса всех вершин которого содержат в себе q :

$$\exists P: u \mapsto v \forall p \in P q \subseteq W_p.$$

Определение 2. Назовем граф $G = \langle V, E \rangle$ *графом смежности*, если любая его пара вершин, пересечение весов которых не пусто, магистрально связна.

Заметим, что полный граф, построенный на множестве вершин V , всегда является графом смежности: любой паре вершин u, v можно сопоставить путь $P = \{u, v\}$.

Определение 3. Граф смежности $G = \langle V, E \rangle$ назовем *нередуцируемым графом смежности*, если после удаления любого ребра e из E , граф $G' = \langle V, E \setminus \{e\} \rangle$ не является графом смежности.

Пусть вершины $u, v \in V$ соединены ребром e .

Определение 4. Нагрузкой ребра e назовем множество $W_e = W_u \cap W_v$.

Определение 5. Множество нагрузок ребер полного графа $G = \langle V, E \rangle$ назовем *универсальным множеством нагрузок* Ω .

Рассмотрим некоторый произвольный вес q .

Определение 6. Сужением графа $G = \langle V, E \rangle$ на вес q назовем граф $G_{/q} = \langle V_q, E_q \rangle$, где

$$V_q = \{v \in V \mid q \subseteq W_v\},$$

$$E_q = \{(uv) \in E \mid u, v \in V_q\}.$$

Вспомогательные факты

В работе [7] было сформулировано утверждение: граф $G = \langle V, E \rangle$ – ГС тогда и только тогда, когда $\forall q \in \Omega$ $G_{/q}$ – связан. Несложно заметить, что G – НГС тогда и только тогда, когда для любого ребра $e \in E$ существует такое $q \in \Omega$, что сужение графа $G^- = \langle V, E \setminus \{e\} \rangle$ на q несвязно.

Утверждение 1. Пусть G – ГС. G – НГС тогда и только тогда, когда для $\forall e \in E$ сужение графа $G^- = \langle V, E \setminus \{e\} \rangle$ на $q = W_e$ несвязно.

Доказательство. Пусть $e = (s, t)$. Так как G – НГС, то существуют u и v такие, что все магистральные пути $P: u \mapsto v$ содержат ребро e . В противном случае e можно исключить из множества ребер, не нарушая магистральной связности графа.

Значит, между u и v не существует магистрального пути в G^- , или, если быть точнее, не существует магистрального пути между вершинами s и t . Значит, не существует пути в сужении графа G^- на вес $W_s \cap W_t = q$, что и требовалось доказать.

Обратное следует напрямую из утверждения, сформулированного в работе [1].

Следствие 1. Из данного утверждения можно сделать вывод: G является НГС тогда и только тогда, когда всякое ребро e из G в сужении на вес W_e является мостом.

Матроиды и графы смежности

Рассмотрим пару $M = \langle E, I \rangle$, где E – произвольное множество, а $I \subseteq 2^E$.

Определение 7. Пару $M = \langle E, I \rangle$ назовем матроидом [10], если она удовлетворяет следующим аксиомам:

- $\emptyset \in I$;
- $A \subseteq B, B \in I \Rightarrow A \in I$;
- $A, B \in I, |A| < |B| \Rightarrow \exists e \in B \setminus A: A \cup \{e\} \in I$.

Множество E в таком случае называют *носителем*, множества из I – *независимыми*, а само семейство I – *семейством независимых множеств*.

Назовем максимальные по включению независимые множества *базами* матроида. Семейство всех баз обозначим через U .

Рассмотрим полный граф $G_{\max} = \langle V, E \rangle$. В качестве носителя матроида выберем множество ребер E . Множество A назовем независимым ($A \in I$), если граф $G_A = \langle V, E \setminus A \rangle$ является графом смежности.

В работе [10] (см. теорему 1.2.3) доказано, что если семейство баз не пусто и $\forall A, B \in U \forall a \in A \setminus B \exists b \in B: A \cup \{b\} \setminus \{a\} \in U$, то $M = \langle E, I \rangle$ – матроид. Покажем, что $M = \langle E, I \rangle$ – матроид.

1) U – непусто. Для начала покажем, что $\emptyset \in I$. Рассмотрим граф $G_f = \langle V, E \rangle$. Данный граф полный, значит, является ГС. Так как $\emptyset \in I$, любое множество $A \in I$ ограничено сверху по включению множеством E , то существует максимальное по включению множество из I , т.е. U непусто.

2) $A, B \in U \forall a \in A \setminus B \exists b \in B : A \cup \{b\} \setminus \{a\} \in U$. Рассмотрим НГС $G_A = \langle V, E_A \rangle$ и $G_B = \langle V, E_B \rangle$, где $E_A = E \setminus A, E_B = E \setminus B$. По сути, требуется показать, что $\forall a \in E_B \setminus E_A \exists b \notin E_B : G = \langle V, E_A \cup \{a\} \setminus \{b\} \rangle$ – НГС.

Возьмем произвольное ребро $a = (s, t) \in E_B \setminus E_A$ веса q и добавим его в граф G_A . Рассмотрим полученный граф $G_A^+ = \langle V, E_A \cup \{a\} \rangle$. Покажем, что любой магистральный путь между вершинами s и t содержит в себе ребра веса q .

Пусть в сужении $G_{A/q}$ существует простой путь P из s в t , который не содержит в себе ребер веса q . Рассмотрим последовательно каждую пару вершин (p_i, p_{i+1}) из пути P . По предположению, $q \subset W_{(p_i, p_{i+1})}$. Так как G_B – НГС, то существует магистральный путь P_i в G_B , соединяющий вершины p_i, p_{i+1} . Соединив последовательно все пути P_i в единую цепочку, мы получим магистральный путь в G_B , не содержащий ребер веса q . Так как нагрузка пути содержит в себе вес q , то каждая вершина будет лежать в сужении $G_{B/q}$. Значит, ребро a не является мостом в $G_{B/q}$, т.е. G_B – не НГС по следствию 1. Получили противоречие. Значит, любой простой путь P из s в t графа G_A содержит хотя бы одно ребро $b = (s', t')$ веса q . Отметим, что по следствию 1 в сужении $G_{A/q}$ ребро b является мостом. Таким образом, любой простой путь между вершинами s и t должен содержать в себе b .

Если удалить из графа G_A^+ такое ребро b , то в полученном графе $G = \langle V, E_A \cup \{a\} \setminus \{b\} \rangle$ вершины s' и t' останутся связанными в сужении $G_{/q}$ (или же магистрально связанными в графе G). Таким образом, G будет ГС.

Покажем, что G – НГС. Удалим ребро a из графа G . Получим граф $G^- = \langle V, E_A \setminus \{b\} \rangle$. Напомним, что b являлось мостом в сужении графа $G_A = \langle V, E_A \rangle$ на собственный вес q . Значит, сужение $G_{/q}^-$ несвязно, т.е. G^- не является ГС.

Удалим произвольное ребро $l \neq a$ из графа G веса r . Получим граф $H^- = \langle V, E_A \cup \{a\} \setminus \{b, l\} \rangle$. В сужении $G_{A/r}$ ребро l являлось мостом. Пусть l соединяло компоненты связности K_x и K_y через вершины x и y соответственно. Если в сужении $G_{/r}$ ребро l – мост, то сужение $H_{/r}^-$ является несвязным. В противном случае в сужение $G_{/r}$ должны были быть добавлены ребра, соединяющие K_x и K_y . Такое ребро всего одно – ребро a .

Так как в $G_{A/r}$ ребро l являлось мостом, то пересечение весов любых двух вершин K_x и K_y равно r . Отсюда получаем, что $W_a = q = r = W_l$. Но тогда в сужении $G_{/r} = G_{/q}$ существует простой путь между вершинами s и t , не проходящий через ребро a . Значит, в сужении $G_{A/r}$ существует простой путь между вершинами s и t , не затрагивающий ребро b , что противоречит рассуждениям выше. Таким образом, любое ребро построенного графа G в сужении на свой вес является мостом, т.е. граф G – НГС.

Из доказанных утверждений следует, что двойка $M = \langle E, I \rangle$ – матроид.

Следствия

В работе [10] (см. лемму 1.2.1) доказано, что любые две базы содержат в себе равное число элементов. Это число называется рангом матроида $r(M)$. Переходя к языку графов смежности, получаем, что число ребер во всех нередуцируемых графах одинаково. Значит, минимальный и нередуцируемый графы суть одно и то же.

Число ребер в минимально графе смежности $G_{\min} = \langle V, E_{\min} \rangle$ можно охарактеризовать числом $|E_{\max}| - r(M)$, где $|E_{\max}|$ – число ребер в полном графе смежности, а $r(M)$ – ранг соответствующего матроида.

Из следствия 1 можно также сделать вывод, что число ребер в минимальном графе $G_{\min}$ есть число мостов во всех сужениях на $q \in \Omega$. Так как всякое сужение $G_{\min}$ связно, то число мостов в любом сужении $G_{\min/q}$ ровно на единицу меньше числа компонент связности в соответствующем графе $G_q = \langle V_q, E_q \rangle$, где

$$V_q = \{v \in V \mid q \subseteq W_v\},$$

$$E_q = \{(u, v) \in E \mid u \in V, v \in V, q \subset W_{(u, v)}\}.$$

Из алгебраической теории графов [11] следует, что число компонент связности в графе G_q есть $|V_q| - \text{rank } A_q$, где A_q – матрица инцидентности графа G_q . Отсюда

$$|E_{\min}| = \sum_{q \in \Omega} [|V_q| - \text{rank } A_q - 1].$$

Заключение

Установлено, что семейство графов смежности над заданной совокупностью фрагментов знаний может быть охарактеризовано с помощью матроида специального вида. Непосредственным следствием этого утверждения является то, что минимальный (по числу ребер) граф смежности является нередуцируемым, а нередуцируемый граф смежности – минимальным. Кроме того, возможность представления семейства графов смежности в виде матроида обеспечивает, в свою очередь, возможность построения минимального графа смежности с помощью жадного алгоритма. При этом достаточно построить один минимальный граф смежности, чтобы определить, возможно ли построение ациклической вторичной структуры алгебраической байесовской сети над заданной совокупностью фрагментов знаний или нет. Если такой минимальный граф смежности будет содержать цикл, то и во всех остальных графах смежности будут обнаруживаться циклы; в этом случае изначальную совокупность фрагментов знаний потребуется модифицировать. Если же хоть один минимальный граф смежности не содержит циклов, то и все остальные минимальные графы смежности будут ациклическими. Наконец, удалось дать точную оценку числу ребер в минимальном графе смежности: оно равно разности числа ребер в максимальном графе смежности и ранга матроида.

Ряд вопросов программной реализации, оценок сложности и корректности алгоритмов построения вторичной структуры АБС был рассмотрен в работах [2, 4, 7–9].

Решенные задачи актуальны не только для теории автоматического обучения алгебраических байесовских сетей, но и для других математических моделей баз фрагментов знаний, где формирование глобальной структуры подчиняется тем же требованиям.

Работа выполнена при финансовой поддержке РФФИ, проект № 09-01-00861-а «Методология построения интеллектуальных систем поддержки принятия решений на основе баз фрагментов знаний с вероятностной неопределенностью».

Литература

1. Опарин В.В., Тулупьев А.Л. Синтез графа смежности с минимальным числом ребер: формализация алгоритма и анализ его корректности // Труды СПИИРАН. – 2009. – Вып. 11. – СПб: Наука, 2009. – С. 142–157.
2. Тулупьев А.Л. Алгебраические байесовские сети: глобальный логико-вероятностный вывод в деревьях смежности: Учеб. пособие. – СПб: СПбГУ; ООО Издательство «Анатолия», 2007. – 40 с.
3. Тулупьев А.Л. Байесовские сети: логико-вероятностный вывод в циклах. – СПб: СПбГУ, 2008. – 140 с.
4. Тулупьев А.Л. Задача локального автоматического обучения в алгебраических байесовских сетях: логико-вероятностный подход // Труды СПИИРАН. – 2008. – Вып. 7. – СПб: Наука, 2008. – С. 11–25.
5. Тулупьев А.Л., Николенко С.И., Сироткин А.В. Байесовские сети: логико-вероятностный подход. – СПб: Наука, 2006. – 607 с.
6. Тулупьев А.Л., Сироткин А.В., Николенко С.И. Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. – СПб: СПбГУ, 2009. – 400 с.
7. Тулупьев А.Л., Столяров Д.М., Ментюков М.В. Представление локальной и глобальной структуры алгебраической байесовской сети в Java-приложениях // Труды СПИИРАН. – 2007. – Вып. 5. – СПб: Наука, 2007. – С. 71–99.
8. Фильченков А.А., Тулупьев А.Л. Структурный анализ систем минимальных графов смежности // Труды СПИИРАН. – 2009. – Вып. 10. – СПб: Наука, 2009. – С. 104–127.
9. Korb K.B., Nicholson A.E. Bayesian Artificial Intelligence. – New York: Chapman and Hall/CRC, 2004. – 364 p.
10. Oxley J.G. Matroid theory. – New York: Oxford University Press, 2006. – 532 p.
11. Norman B. Algebraic graph theory (2nd edition). – Cambridge: Cambridge University Press, 1996. – 205 p.

Опарин Всеволод Владиславович – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, студент, oparin.vsevolod@gmail.com
Фильченков Андрей Александрович – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, студент, aaafil@mail.ru
Сироткин Александр Владимирович – Санкт-Петербургский институт информатики и автоматизации РАН, младший научный сотрудник, avs@iiias.spb.su
Тулупьев Александр Львович – Санкт-Петербургский институт информатики и автоматизации РАН, доктор физ.-мат. наук, доцент, зав. лабораторией, ALT@iiias.spb.su, ALT4488@peterstar.ru