Проверка ацикличности алгебраической байесовской сети с применением третичной структуры

Вяткин Артём Андреевич, студент математико-механического факультета СПбГУ, vyatkin.artex@gmail.com
Харитонов Никита Алексеевич, аспирант СПбГУ, nak@dscs.pro
Тулупьев Александр Львович, проф. каф. инф., д-р физ.-мат. наук, проф., alt@dscs.pro

Аннотация

В теории алгебраических байесовских сетей (АБС) существует понятие ацикличности алгебраической байесовской сети. Это свойство имеет место быть, когда АБС представима в виде дерева смежности. В данной работе приведен способ проверки непротиворечивости АБС с применением третичной структуры, что может быть полезно при независимом использовании третичной структуры, например, при глобальном апостериорном выводе.

Введение

Зачастую необходимо опираться на знания, собранные в данной конкретной области, для того, чтобы составлять планы будущей деятельности, находить решения, размышлять в терминах этой предметной области. Подобные знания большей своей частью заключают в себе неопределенность, анализ и обработка которых является одной из сторон современной информатики [6]. Такие знания могут рассматриваться в виде совокупности утверждений, заключающих между собой логические и стохастические связи. Данные соотношения между утверждениями способны охарактеризовать эксперты рассматриваемой области. В то же время суждения экспертов характеризуют связи между небольшим набором сущностей из предметной области. В итоге комплекс знаний экспертов разделяется на отдельные компоненты, фрагменты знаний, которые в общей сложности образуют базу фрагментов знаний [5].

Для описания подобной экспертной системы необходима математическая модель, в качестве которой могут выступать вероятностные графические модели (ВГМ). Вероятностные графические модели находят свое применение в разных задачах, таких, как, например, исследование влияния окружающей среды и генов на получаемые заболевания [3], анализ социоинженерных атак [2]. В класс вероятностных графических моделей также входят алгебраические байесовские сети. В качестве математической модели фрагмента знаний в теории алгебраических байесовских сетей могут выступать идеалы конъюнктов, определенных над алфавитом из набора пропозициональных переменных. Каждому элементу идеала приписывается либо точечная, либо интервальная оценка вероятности истинности. В совокупности все фрагменты знаний составляют алгебраическую байесовскую сеть.

Алгебраическая байесовская сеть может рассматриваться с использованием различных типов структур. Первичная структура представляет из себя обычное множество фрагментов знаний, без указания связей между ними. Вторичная структура, как и третичная, представляются в виде специального рода графов, использующих фрагменты знаний и их пересечения в качестве нагрузок ребер и вершин. Вторичная структура является графом смежности, при этом для одной заданной первичной структуры может быть несколько различных вторичных структур. Важным понятием является непротиворечивость алгебраической байесовской сети — согласованность оценок вероятности истинности элементов фрагментов знаний. Еще одним важным свойством первичной структуры и алгебраической байесовской сети в целом служит их ацикличность — возможность построить вторичную структуру, являющуюся деревом смежности. Наличие ацикличности у алгебраической байесовской сети позволит уменьшить вычислительную сложность проверки и поддержания непротиворечивости. Также именно для ациклических байесовских сетей доказана корректность применения алгоритма глобального апостериорного вывода, основанного на распространении виртуального свидетельства и заключающегося в переоценке вероятности истинности элементов фрагментов знаний на основе новых поступивших знаний — свидетельств. При этом для пропагации свидетельств может быть использована только третичная структура, применяющаяся сейчас для построения вторичных структур, используемых в дальнейшем для распространения свидетельств. Таким образом, проверка ацикличности с применением третичной структуры будет полезна при ее обособленном от вторичной структуры использовании и данная работа посвящена решению этого вопроса.

Теоретическая основа

В данной главе опишем систему терминов и ряд алгоритмов, используемых в работе и приведенных в [4, 7, 8].

Основные определения

Прежде всего рассмотрим объекты, которые будут соответствовать переменным, заключающим утверждения. Они образуют *алфавит* — множество, состоящее из атомарных пропозициональных формул (которые могут называться атомами). $A = \{x_1, ..., x_n\}$ определяет алфавит из n атомов. Для оценки самих атомарных пропозиций, а также связей между ними определим *идеал конъюнктов*, построенный над алфавитом $A = \{x_1, ..., x_n\}$ — множество формул вида $\{x_{i_1}x_{i_2}...x_{i_k} \mid 0 \leqslant i_1 < ... < i_k \leqslant n-1, k \leqslant n\}$, где $x_{i_1}x_{i_2}...x_{i_k}$ представляет конъюнкцию соответствующих переменных.

Фрагментом знаний (математической моделью фрагмента знаний), который построен над алфавитом A, назовем пару (C, p), где C — идеал конъюнктов над соответствующим алфавитом, p — интервальные или скалярные (точечные) оценки вероятностей для каждого конъюнкта из идеала C.

Для дальнейшей работы с фрагментами знаний и их наборами удобно определить вес, который соответствует каждому ФЗ — нагрузкой или весом фрагмента знаний $W(C, p)$ назовем подалфавит алфавита, над которым задан фрагмент знаний, $W(C, p) = \{x_i \mid x_i \in C, x_i \in A\}$.

Назовем *набором максимальных фрагментов знаний* (набор МФЗ, первичная структура алгебраической байесовской сети) такой набор фрагментов знаний, что никакая нагрузка фрагмента знаний не содержится полностью в нагрузке другого фрагмента знаний из представленного набора. То есть $\forall i \neq j$ выполнено: $W(V_i) \not\subset W(V_j)$ и $W(V_j) \not\subset W(V_i)$.

Вторичная структура АБС

Сепаратором двух МФЗ, V_i и V_j, назовем подалфавит, который является пересечением нагрузок этих ФЗ: $W(V_i, V_j) = W(V_i) \cap W(V_j), i \neq j$. Пара МФЗ называются сочлененными, если их сепаратор непуст.

Граф максимальных фрагментов знаний — ненаправленный граф, вершинам которого сопоставлены МФЗ, вошедшие в АБС и ребра возможны только между сочлененными ФЗ. *Нагрузкой* $W(\{V_i, V_j\})$ *ребра* $\{V_i, V_j\} \in E(G)$ графа G назовем сепаратор его концов: $W(\{V_i, V_j\}) = W(V_i) \cap W(V_j)$. Определим и *нагрузку* $W(H)$ *подграфа* $H \subseteq G$ — наибольший по включения подалфавит, входящий в нагрузку всех вершин подграфа: $W(H) = \bigcap_{V \in H} W(V)$.

Магистральный путь между сочлененными вершинами V_i и V_j — такой путь между этими вершинами, что нагрузка каждой вершины пути содержит сепаратор концов этого пути. Далее граф будет *магистрально связен*, если между каждой из его сочлененных вершин существует магистральный путь.

Граф смежности — магистрально связный граф МФЗ. *Дерево смежности* — граф смежности, представимый в виде дерева.

В результате, помимо первичной структуры АБС, можно дать определение *вторичной*. Такой структурой будет являться некоторый граф смежности АБС.

Так же существует понятие *максимального графа смежности* G_{max} — наибольшего по числу ребер графа смежности. Для заданного множества вершин существует единственный максимальный граф смежности, то есть тот, в котором между вершинами существует ребро только тогда, когда они сочлененные.

Дополнительно предположим, что первичная структура *связна*, то есть связен максимальный граф смежности, построенный над этой структурой. В противном случае можно рассматривать наборы вершин из каждой компоненты связности как отдельные АБС.

Третичная структура АБС

Сужением $G \downarrow U$ графа G на нагрузку U назовем граф, в который входят те и только те ребра и вершины исходного графа G, нагрузки которых равны или содержат U. *Значимое сужение* — сужение на нагрузку, которая является сепаратором для некоторой пары МФЗ. На сужение можно наложить дополнительные ограничения, тогда получим *сильное сужение* $G \twoheaddownarrow U$ — сужение $G \downarrow U$, из которого удалили все ребра нагрузки U. После сильного сужения граф $G \downarrow U$ разбивается на компоненты связности, после сужения же $G \downarrow U$ граф остается связным.

Одним из основных объектов в новоопределяемой структуре будет *значимая нагрузка* U — непустой сепаратор некоторой пары ФЗ первичной структуры. *Замкнутым же снизу множеством нагрузок* назовем объединение множества значимых нагрузок с множеством нагрузок вершин МФЗ. *Замкнутое множество нагрузок* — объединение замкнутого снизу множество нагрузок с одноэлементным множеством, содержащим пустое множество.

При этом на множестве нагрузок существует частичный порядок, являющийся отношением включения. Таким образом, *родительским графом* (третичной структурой АБС) назовем диаграмму Хассе замкнутого множества нагрузок. Диаграмму Хассе можно рассматривать как транзитивное сокращение, поэтому родительский граф единственный при заданной первичной структуре АБС [1].

Проверка ацикличности

Алгоритм проверки того, что вторичная структура представима в виде дерева смежности основывается на следующей теореме, описанной в [9]:

Теорема 1. *Связная первичная структура АБС циклична тогда и только тогда, когда не выполняется соотношение:*

$$|MKP| = \sum_{U \in Sep} Conn(G \downarrow U) - |Sep| + 1$$

где MKP — первичная структура АБС, набор ФЗ, $Conn(G \downarrow U)$ — число компонент связности графа $G \downarrow U$, Sep — множество непустых сепараторов.

Все слагаемые из выражения данной теоремы можно подсчитать, используя только первичную и вторичную структуры АБС, при этом наибольшую сложность здесь представляет расчет числа компонент связности, количество сепараторов же равно количеству вершин родительского графа, за исключением верхней, пустой вершины, и листьев-фрагментов знаний. Поэтому далее рассмотрим, как можно подсчитать сумму компонент связности графов сильных сужений с использованием третичной структуры. Для этого докажем теорему:

Теорема 2. *Две вершины с нагрузками в виде фрагментов знаний kp_1 и kp_2 лежат в одной компоненте связности C_u графа $G \downarrow u$, полученной после сильного сужения на значимую нагрузку u, тогда и только тогда, когда существует последовательность таких вершин с нагрузками в виде фрагментов знаний и с kp_1 и kp_2 как крайними элементами, что для каждых двух соседних вершин в этой последовательности существует нагрузка, которая является предком по отношению к этим вершинам и потомком по отношению к вершине с нагрузкой u в родительском графе. Формально:*

$$kp_1, kp_2 \in C_u \Leftrightarrow \exists v_1 = kp_1, ..., v_n = kp_2 :$$
$$\forall i = 1, ..., n-1 : (\exists w_i : kp_1, kp_2 \in descendants(w_i) \;\&\; w_i \in descendants(u))$$

Доказательство. Действительно, если kp_1 и kp_2 лежат в одной компоненте связности, то между ними существует путь, связанный ребрами, нагрузки которых включают, но не равны нагрузке сильного сужения u. С другой стороны, такая последовательность и будет путем, связывающим kp_1 и kp_2 в графе $G \downarrow u$, ведь наличие общей вершины-предка между двумя соседними элементами последовательности означает то, что ребро между этими элементами будет содержать вес, включающий вес вершины-предка. Но так как в последовательности подобные вершины-предки являются потомками по отношению к u, то вес связующих ребер будет включать, но не равняться u, что и означает наличие пути в $G \downarrow u$. ■

Замечание 1. Отметим, что если в родительском графе между двумя вершинами существует общая вершина-потомок, то все фрагменты знаний, являющиеся потомками по отношению к первым двум вершинам будут лежать в одной компоненте связности, полученной после сильного сужения на общую для этих двух вершин нагрузку вершины-предка. То есть, если $\exists w_1, w_2 :$ $kp_1 \in \text{descendants}(w_1)$ & $kp_2 \in \text{descendants}(w_2)$ & $w_1, w_2 \in \text{descendants}(u)$, а также $\exists w_3 : w_3 \in \text{descendants}(w_1) \cap \text{descendants}(w_2)$, то $kp_1, kp_2 \in C_u$.

Основываясь на доказанной теореме, предложим следующий алгоритм подсчета числа компонент связности. Пусть нам необходимо найти количество компонент связности для вершины u. Тогда распространим по каждой дочерней вершине u различные маркеры, назовем их *цветами*. Затем от каждой такой вершины будем распространять по дочерним узлам цвет, полученный ранее. Если в одну вершину поступило несколько разных цветов, то признаем эти цвета одинаковыми. В итоге, количество цветов, оставшихся после распространения их до листьев и будет совпадать с количеством компонент связности $G \downarrow u$.

Алгоритм работает корректно. Если две вершины с фрагментами знаний в качестве нагрузок получили один цвет, то, по замечанию 1, они будут лежать в одной компоненте связности. С другой стороны, предположим, что два листа kp_1 и kp_2 в родительском графе получили разные цвета, но лежат в одной компоненте. Тогда, по утверждению 2, будут существовать последовательности из вершин $v_1, ..., v_n$ и $w_1, ..., w_{n-1}$, которые по действию алгоритма должны быть окрашены в один цвет. Но в таком случае $v_1 = kp_1$ и $v_n = kp_n$ будут окрашены в единый цвет — противоречие.

Заключение

В данной работе были представлен алгоритм, позволяющий применять только третичную структуру АБС для проверки ацикличности, а также доказана его корректность. Применение этого алгоритма может быть полезно при обособленном использовании третичной структуры, например при ее использовании для глобального апостериорного вывода.

Список литературы

[1] Aho A., Garey M., Ullman J. The Transitive Reduction of a Directed Graph // SIAM Journal on Computing. 1972. Vol. 1, No. 2. P. 131–137.

[2] Khlobystova A. O., Abramov M. V., Tulupyev A.L. An approach to estimating of criticality of social engineering attacks traces // Studies in Systems, Decision and Control. 2019. Vol. 199. P. 446–456.

[3] Su C., Andrew A., Karagas M.R. et al. Using Bayesian networks to discover relations between genes, environment, and disease // BioData Mining. 2013. Vol. 6, No. 6.

[4] Сироткин А. В., Тулупьев А. Л. Моделирование знаний и рассуждений в условиях неопределенности: матрично-векторная формализация локального синтеза согласованных оценок истинности // Труды СПИИРАН. 2011. Вып. 18. С. 108–135.

[5] Тулупьев А. Л. Алгебраические байесовские сети: локальный логико-вероятностный вывод: Учеб. пособие. // СПб.: ООО Издательство «Анатолия», 2007. 80 с.

[6] Тулупьев А. Л., Николенко С. И., Сироткин А. В. Байесовские сети: логико-вероятностный подход. // СПб.: Наука, 2006. 607 с.

[7] Тулупьев А. Л., Сироткин А. В. Локальный апостериорный вывод в алгебраических байесовских сетях как система матрично-векторных операций // Интегрированные модели и мягкие вычисления в искусственном интеллекте. V-я Международная научно-практическая конференция, 9 сентября — 12 сентября 2009 г. Сборник научных трудов. В 2-х т. Т. 1. СПб.: Наука, 2009. С. 425–434.

[8] Фильченков А. А., Тулупьев А. Л. Третичная структура алгебраической байесовской сети // Труды СПИИРАН. 2011. Вып. 18. С. 164–187.

[9] Фильченков А. А., Тулупьев А.Л. Связность и ацикличность первичной структуры алгебраической байесовской сети // Вестник Санкт-Петербургского государственного университета. Серия 1. Математика. Механика. Астрономия. 2013. Вып. 1. С. 110–119.