Алгебраические байесовские сети: нахождение канонического представителя фрагмента знаний методом Монте-Карло

Н. А. Харитонов
Санкт-Петербургский государственный университет
nak@dscs.pro

А. Л. Тулупьев
Санкт-Петербургский государственный университет;
Санкт-Петербургский федеральный исследовательский центр Российской академии наук
alt@dscs.pro

***Аннотация*. Алгебраические байесовские сети относятся к классу вероятностных графических моделей. Одним из базисов, на которых построена теория алгебраических байесовских сетей, является декомпозиция знаний на фрагменты знаний. Фрагмент знаний может быть представлен в виде набора пропозиций-квантов с сопоставленными им интервальными или скалярными оценками вероятности истинности. При этом операции логико-вероятностного вывода во фрагменте знаний с интервальными оценками имеют значительно большую сложность по сравнению с фрагментом знаний со скалярными оценками. Таким образом, видится важной задача поиска фрагмента знаний со скалярными оценками, наиболее полно отражающего информацию о фрагменте знаний с интервальными, или, иными словами, канонического представителя фрагмента знаний с интервальными оценками элементов.**

Решение поставленной задачи сводится к поиску канонического представителя фрагмента знаний методом Монте-Карло. Приведены теоретические обоснования использования данного метода, алгоритм нахождения, результаты экспериментов для заданного фрагмента знаний и методология применения метода на всей алгебраической байесовской сети.

Ключевые слова: алгебраические байесовские сети, вероятностные графические модели, фрагмент знаний, метод Монте-Карло, канонический представитель

I. Введение

Алгебраические байесовские сети относятся к классу вероятностных графических моделей. Идеей, лежащей в основе их концепции, является декомпозиция знаний на малые части, называемые фрагментами знаний. Одним из классических представлений последних является набор квантов с сопоставленными им скалярными или интервальными оценками вероятности истинности.

Возникающие в ходе обучения и работы с алгебраическими байесовскими сетями интервальные оценки удобны при интерпретации знаний эксперта, однако сложны с алгоритмической точки зрения.

Работа выполнена в рамках проекта по государственному заданию СПб ФИЦ РАН № 0073-2019-0003 и поддержана Санкт-Петербургским государственным университетом, проект № 73555239.

Таким образом, в ряде случаев может возникнуть ситуация, когда вместо интервальных оценок требуется обрабатывать скалярные.

Данная работа посвящена получению фрагмента знаний со скалярными оценками на основе имеющегося фрагмента знаний с интервальными оценками вероятности истинности элемента методом Монте-Карло.

II. Актуальность и методология

Вероятностные графические модели, в частности, байесовские сети доверия, родственные алгебраическим байесовским сетям, находят практическое применение в инженерии, медицине и прочих областях науки, требующих прогнозирования и анализа рисков [7, 8, 9].

В контексте алгебраических байесовских сетей используется система терминов, теорем и обозначений, определенная в работах [11, 12].

Алгебраические байесовские сети представляют собой модель фрагментов без знаний, связи между которыми могут быть представлены в виде графа. Каждый из фрагментов знаний в свою очередь представляет собой идеал дизъюнктов или конъюнктов или набор пропозиций-квантов (в дальнейшем слово «пропозиций» будет опускаться), каждому из которых сопоставлена интервальная оценка вероятности. Представление сети в том или ином виде целостно, то есть, алгебраическая байесовская сеть не может одновременно содержать в себе фрагмент знаний, представленный в виде идеала дизъюнктов и фрагмент знаний, представленный в виде идеала конъюнктов или набора квантов. При этом каждая из форм представления может быть преобразована в другую, для чего определяется целое семейство наборов матриц переходов.

Под квантом здесь и ниже понимается конъюнкция переменных заданного алфавита и их отрицаний. Набор квантов – набор всевозможных квантов заданной длины. Например, для алфавита $\{x_1, x_2\}$ набор квантов будет выглядеть следующим образом:

$$Q = \{q_0, q_1, q_2, q_3\} = \{\underline{x}_1\underline{x}_2, \underline{x}_1 x_2, x_1\underline{x}_2, x_1 x_2\}.$$

Базовыми операциями логико-вероятностного вывода над алгебраическими байесовскими сетями являются априорный и апостериорный выводы и поддержание непротиворечивости. Приведенные операции делятся на локальные, происходящие в рамках отдельно взятого фрагмента знаний, и глобальные, происходящие во всей сети.

При обучении алгебраических байесовских сетей на входных данных получаемые фрагменты знаний содержат в себе набор квантов, имеющий интервальные оценки вероятностей [4].

Интервальные оценки удобны тем, что они отражают неполноту, неточность, нечеткость содержащейся в сети информации, что позволяет, в том числе, обрабатывать экспертные оценки, представленные в виде неточных лексических конструкций, не прибегая к искусственному уточнению оценки. Например, фраза «около трети» может быть интерпретирована как интервал [0.28;0.37], а «почти половина» как [0.4;0.6].

При этом операции логико-вероятностного вывода над сетью с интервальными оценками имеют большую сложность, чем над сетью со скалярными оценками. Так, поддержание непротиворечивости во фрагменте знаний с интервальными оценками подразумевает под собой решение задачи линейного программирования, в то время как во фрагменте знаний со скалярными оценками происходит лишь проверка наличия решений в матричном уравнении.

Таким образом, видится актуальной задача нахождения канонического представителя фрагмента знаний с интервальными оценками, представленного в виде фрагмента знаний со скалярными оценками.

Заметим, что фрагмент знаний с интервальными оценками можно рассматривать как некоторую фигуру в n-мерном пространстве. При этом канонически представитель будет не чем иным, как одной из точек, а в идеальном случае – центром масс данной фигуры.

Одним из возможных решений данной задачи видится использование метода Монте-Карло [1, 3]. При этом можно рассматривать не собственно область, ограниченную фрагментом знаний, но симплекс, содержащий в себе эту область.

III. Создание канонического представителя

Для получения канонического представителя мы рассматриваем фрагмент знаний с интервальными оценками KP_{int}, представленный в виде набора квантов. Каждый квант является независимым событием, поэтому сумма вероятностей квантов во фрагменте знаний со скалярными оценками равна единице.

Таким образом, перед нами встает задача нахождения таких квантов $q_0, q_1, \dots q_n$, что

$$p(q_0) + p(q_1) + \dots + p(q_3) = 1.$$

Иначе говоря, речь идет о стандартном n-мерном симплексе [10]:

$$\Delta^n = \{(x_0, \dots x_n): \sum_0^n \quad x_i = 1;\ x_i \geq 0 \vee i\}.$$

Однако есть небольшая оговорка: все рассматриваемые кванты должны принадлежать множеству области, ограничиваемой KP_{int}. В дальнейшем, не умаляя общности, будем отождествлять эту область с самим фрагментом знаний.

Этот факт учитывается при создании канонического представителя следующим образом: из множества полученных точек симплекса необходимо отсекать те точки, которые относятся к симплексу, но не относятся к KP_{int}.

Генерация точек симплекса происходит в два шага:

- на первом шаге создается точка, значение каждой координаты которой независимо генерируются по закону гамма-распределения;
- на втором шаге координаты полученной точки нормируется путем поочередного деления значений координат на их сумму.

Если точка симплекса не попадает во фрагмент знаний, она генерируется заново.

После определенного числа итераций полученное множество точек строится линейная комбинация точек. Координаты точки, полученной в результате линейной комбинации, являются ничем иным, как набором вероятностей канонического представителя фрагмента знаний.

Кроме поиска собственно канонического представителя, видится важным оценка его «точности», то есть того, насколько информация в представителе полно отражает информацию, представленную в KP_{int}, и, как следствие, косвенная оценка количества информации, отсеченной от имеющейся изначально.

В данной работе для оценки этого параметра будет использоваться дисперсия и среднее квадратичное отклонение, рассчитанные для набора, сгенерированных в процессе получения канонического представителя, точек.

IV. Алгоритм

Алгоритм нахождения канонического представления состоит из двух функций, одна из которых вложена в другую.

«Внутренняя» функция вызывается в цикле и возвращает точку симплекса, входящую в KP_{int}.

Внешняя функция сохраняет полученные точки в наборе и возвращает их усредненное значение, а также дисперсию и среднее квадратичное отклонение набора.

```
var KP          //исходный фрагмент знаний
int N //число итераций
// Внутренняя функция
func FindScalarKP(KP, N):
    var points[N]       //набор точек
    for i in N:
        // генерация точки
        tmp = generatePoint(KP)
        // добавление в набор
        kps.add(tmp)
    res := points/len(points)
    // нахождение дисперсии и с.к.о.
    d := dispersion(points)
    sko := sko(points)
    return res, d, sko

// внутренняя функция
func generatePoint(KP):
    // точка - массив координат размерности
    // фрагмента знаний
    point[len(KP)]
    do
        point = new[len(KP)]
        for i in range len(KP):
            // получение координаты
            // из гамма-распределения
            coord := gamma(2,2)
            point.add(gamma)
        // деление всех элементов массива
        // на их сумму (нормализация)
        point = point/sum(point)
    // условие цикла - точка находится
    // во фрагменте знаний, иными
    // словами каждая координата между
    // верхней и нижней границами
    // оценки фрагмента знаний
    while in(point, KP)
    point = point/sum(point)
    return point
```

V. Пример

Рассмотрим фрагмент знаний, являющийся набором из четырех квантов.

$$p(q_0) = [0.2; 0.55],$$
$$p(q_1) = [0.05; 0.4],$$
$$p(q_2) = [0.3; 0.65],$$
$$p(q_3) = [0.1; 0.45].$$

Будем рассматривать итерации при N равном 100, 1000 и 10000.

В приведенной ниже таблицы представлены результаты: полученная точка, дисперсия и среднее квадратичное отклонение.

N	Канон. пр.	Дисперсия	С.к.о.
100	$p(q_0) = 0.284$ $p(q_1) = 0.141$ $p(q_2) = 0.381$ $p(q_3) = 0.193$	0.125	0.354
1000	$p(q_0) = 0.283$ $p(q_1) = 0.145$ $p(q_2) = 0.380$ $p(q_3) = 0.192$	0.124	0.352
10000	$p(q_0) = 0.285$ $p(q_1) = 0.144$ $p(q_2) = 0.383$ $p(q_3) = 0.189$	0.124	0.353

Таким образом, канонические представители, полученные при 100, 1000 и 10000 итераций имеют близкие дисперсию и стандартное квадратичное отклонение, что позволяет использовать 100–1000 итераций в алгоритме.

VI. Канонический представитель алгебраической байесовской сети

Отметим, что изучение вопроса получения канонического представителя алгебраической байесовской сети является дальнейшим направлением развития представленной работы.

Здесь же приведем «наивный» способ его получения.

Как уже было отмечено, алгебраическая байесовская сеть может быть представлена в виде набора фрагментов знаний. Условно пронумеруем их от 1 до n.

Канонический представитель для первого фрагмента знаний может быть получен способом, представленным выше.

Далее для каждого из фрагментов знаний, имеющих общие кванты с первым, вероятности квантов на пересечении могут быть представлены как интервалы с совпадающими верхней и нижней границами, равными вероятности кванта из уже построенного канонического представителя. После поддержания непротиворечивости канонические представители для данных фрагментов знаний могут быть поочередно подсчитаны приведенным выше образом с уточнением оценок на каждом шаге.

Однако данный подход при определенных условиях может привести к некорректности полученных оценок (фрагменты знаний буду оказываться противоречивыми). Таким образом, данный вопрос требует дальнейшего изучения.

VII. Заключение

В работе представлен подход к получению канонического представителя методом Монте-Карло. Приведено обоснование применения данного способа, алгоритм получения. Проведено сравнение точности получаемых оценок при разном числе итераций алгоритма. Предложен наивный метод получения канонического представителя алгебраической байесовской сети.

Дальнейшими направлениями исследования являются изучение канонического представителя алгебраической байесовской сети и применение полученных знаний в контексте прикладных исследований, в том числе в области оценки защищенности системы от социоинженерных атак [2, 5, 6].

Список литературы

[1] Altmann Y., McLaughlin S., Dobigeon N. Sampling from a multivariate Gaussian distribution truncated on a simplex: A review // IEEE Workshop on Statistical Signal Processing Proceedings. 2014. Art. 6884588. P 113-116.

[2] Bushmelev F.V., Abramov M.V., Tulupyeva T.V. Adaptive Method of Color Selection in Application to Social Media Images // Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 8th International Conference on “Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT 2020)”. Smolensk, Russia, June 29–July 1, 2020. P. 252-257.

[3] Comas-Cufí M., Martín-Fernández J.A., Mateu-Figueras G., Palarea-Albaladejo J. Modelling count data using the logratio-normal-multinomial distribution // SORT. 2020. Vol. 44, iss 1. P. 99-126.

[4] Kharitonov N.A., Maximov A.G., Tulupyev A.L. Algebraic Bayesian Networks: Naïve Frequentist Approach to Local Machine Learning Based on Imperfect Information from Social Media and Expert Estimates // Communications in Computer and Information Science. 2019. №1093.

[5] Khlobystova A.O, Abramov M.V. The models separation of access rights of users to critical documents of information system as factor of reduce impact of successful social engineering attacks // Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 8th International Conference on “Fuzzy Systems, Soft Computing and Intelligent Technologies (FSSCIT 2020)”. Smolensk, Russia, June 29–July 1, 2020. P. 264–268.

[6] Korepanova A.A., Oliseenko V.D., Abramov M. V. Applicability of Similarity Coefficients in Social Circle Matching // 2020 XXIII International Conference on Soft Computing and Measurements (SCM). St. Petersburg, Russia. 2020. pp. 41-43. doi: 10.1109/SCM50615.2020.9198782.

[7] Liang R., Liu F., Liu J. A belief network reasoning framework for fault localization in communication networks // Sensors (Switzerland). 2020. Vol. 20, iss. 3. Art. 6950. P. 1-21.

[8] Steijn W.M.P., Van Kampen J.N., Van der Beek D., Groeneweg J., Van Gelder P.H.A.J.M. An integration of human factors into quantitative risk analysis using Bayesian Belief Networks towards developing a ‘QRA+’ // Safety Science. 2020. Vol. 122. Art. 104514.

[9] Wu Y., McLeod C., Blyth C., Bowen A., Martin A., Nicholson A., Mascaro S., Snelling T. Predicting the causative pathogen among children with osteomyelitis using Bayesian networks – improving antibiotic selection in clinical practice // Artificial Intelligence in Medicine. 2020. Vol. 107. Art. 101895.

[10] Кострикин А.И., Манин Ю.И. Линейная алгебра и геометрия. 2-е изд. М.: Наука, 1986. 304 с.

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

[12] Тулупьев А.Л., Сироткин А.В., Николенко С.И. Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. СПб.: Изд-во С.-Петерб. ун-та, 2009.