

Алгебраические байесовские сети: использование канонического представителя фрагмента знаний при локальном выводе

В. Е. Лоскутов¹, А. А. Вяткин²

¹Санкт-Петербургский государственный университет

²Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

¹v-loskutov@inbox.ru, ²aav@dscs.pro

Аннотация. Теория вероятностных графических моделей является одной из основ для исследования методов представления и обработки вероятностных зависимостей между переменными, заменяя необходимость хранения всех возможных совместных распределений вероятностей на графическое представление, отражающее условные зависимости. Одним из представителей вероятностных графических моделей являются алгебраические байесовские сети, в которых узлы сети представляют модели фрагмента знаний, а ребра отражают их взаимосвязь. Каждому элементу из фрагмента знаний присваивается оценка вероятности истинности — скалярная или интервальная. В ситуации, когда фрагменты знаний имеют интервальные оценки, обработка может занимать значительное время, что может потребовать поиска канонического представителя фрагмента знаний, то есть перехода от интервальных оценок к скалярным. На текущий момент известно два метода построения канонического представителя: долгий, но точный и быстрый, но приближенный, использующий метод Монте-Карло. Данная работа заключается в определении алгоритма автоматизации построения канонического представителя фрагмента знаний и оценке погрешности, возникающей при использовании приближенного алгоритма в процессе апостериорного вывода — одного из основных способов использования алгебраических байесовских сетей и фрагментов знаний на практике. В ходе работы была найдена оценка точности результатов при использовании приближенного алгоритма для фрагментов знаний, состоящих из 1–3 квантов, которая составила около 0.02 единиц оценки вероятности.

Ключевые слова: алгебраические байесовские сети; канонический представитель; точный метод поиска канонического представителя; метод Монте-Карло; фрагмент знаний; вероятностные графические модели; машинное обучение

I. ВВЕДЕНИЕ

В современных исследованиях в областях машинного обучения вероятностные графические модели играют большую роль и находят применение в таких областях, как изучение распределения транспортных потоков [1], обработка изображений [2], оценка кредитного риска [3], обнаружение мошенничества [4], рекомендательные системы [5]. Одним из классов таких моделей является класс алгебраических байесовских сетей (АБС), позволяющих формализовать знания о предметной

области в виде утверждений с заданной степенью уверенности в их истинности, формализуемой в виде оценок вероятности истинности.

Важной фундаментальной идеей в рамках теории АБС является концепция декомпозируемости знаний на фрагменты знаний (ФЗ). Фрагменты знаний представляют собой небольшие наборы утверждений со связями между ними, которые выступают отдельными частями для формирования глобальной структуры всей АБС. Примером формализации (математической модели) ФЗ служат наборы квантов, для каждого из которых устанавливается скалярная или интервальная оценка вероятности истинности. Далее под ФЗ мы будем иметь в виду именно математическую модель. Использование интервальных оценок предоставляет дополнительные возможности интерпретации, позволяет точнее отражать степень уверенности в истинности утверждений.

Тем не менее, применение интервальных оценок сопряжено с увеличением временных затрат на работу алгоритмов, что может создать проблемы в условиях ограниченных вычислительных ресурсов. Это подчеркивает необходимость в определении эффективного подхода, заключающегося в переходе от ФЗ с интервальными оценками к ФЗ со скалярными. Получение скалярных оценок, максимально полно отражающих информацию, заложенную в интервальных оценках, определяется как построение канонического представителя. Такой подход может уменьшить время на обработку ФЗ, жертвуя, однако, полнотой заключаемой в ФЗ информации. На текущий момент известно два алгоритма построения канонического представителя: точный [6] и приближенный [7].

Применение АБС и ФЗ на практике сопряжено с использованием алгоритмов вывода, одним из которых является алгоритм апостериорного вывода, позволяющий моделировать учет новой информации о предметной области. Целью же данной работы является определение алгоритма автоматизации построения канонического представителя фрагмента знаний, а также оценка погрешности приближенного алгоритма при апостериорном выводе.

Таким образом, теоретическая значимость работы выражается в разработке подхода к автоматизации построения канонического представителя ФЗ. Практическая значимость же заключается в оценке точности результатов при использовании приближенного

алгоритма в апостериорном выводе, что позволит расширить возможности применения АБС для решения различных практических задач.

II. ПОСТАНОВКА ЗАДАЧИ

В данном разделе формализуем ключевые термины, которые будут использоваться в рамках исследования, а также поставим задачу. Описание определений и методов основывается на предыдущих исследованиях в области теории алгебраических байесовских сетей [6–9].

Для начала определим математическую модель фрагмента знаний. Обозначим **атомом** элементарную пропозицию, касающуюся предметной области. В то же время **алфавит** $A = \{x_1, \dots, x_n\}$ представляет собой множество атомов x_i , которые служат основой для формализации системы знаний.

На основе атомов мы можем конструировать **конъюнкты**, представляющие собой конъюнкции, состоящие из атомов. Более того, мы можем говорить о **квантах** — конъюнкциях всех атомов алфавита, а также их отрицаний. Для объединения конъюнктов в единый набор вводится понятие **идеала конъюнктов** C , который включает все возможные конъюнкты, составляемые на основе атомов алфавита A . В свою очередь, **набор квантов** Q включает в себя все возможные кванты, которые формируются из атомов алфавита. Например, для алфавита $\{x_0, x_1\}$ идеалом конъюнктов будет $\{\varepsilon, x_0, x_1, x_0x_1\}$ где ε — пустая конъюнкция, а набором квантов — $\{\bar{x}_1\bar{x}_0, \bar{x}_1x_0, x_1\bar{x}_0, x_1x_0\}$.

Теперь мы можем перейти к формализации математической модели фрагмента знаний (ФЗ) со скалярными оценками, который определяется как пара (C, p) , где:

C представляет идеал конъюнктов; p — это функция, заданная как $p: C \rightarrow [0, 1]$ и называемая скалярной оценкой вероятности истинности.

Данное выше определение для ФЗ — основное, однако также существует представление ФЗ в виде набора квантов, и между оценками вероятности истинности для этих представлений существуют взаимно однозначные матричные переходы.

Помимо этого, ФЗ может иметь интервальные оценки [9]. Согласно построению [6], ФЗ с интервальными оценками можно рассматривать как семейство ФЗ со скалярными оценками. Переход от интервальных оценок к скалярным, при котором объект со скалярными оценками максимально полно представляет объект с интервальными, определим как построение **канонического представителя**. Канонический представитель можно строить как для отдельных фрагментов знаний, так и для всей алгебраической байесовской сети, в своей первичной структуре являющейся набором фрагментов знаний.

На текущий момент описано два алгоритма нахождения канонического представителя для фрагмента знаний — точный [6] и приближенный [7]. Оба этих алгоритма рассматривают интервальный ФЗ, представленный в виде набора квантов с оценками вероятности, как выпуклый многогранник, лежащий в N -

мерном симплексе, где $N = 2^n - 1$, и n — мощность алфавита, над которым построен ФЗ. В таком случае, каждая точка в таком многограннике будет определять ФЗ со скалярными оценками, лежащий в семействе ФЗ, задаваемом ФЗ с интервальными оценками. В силу такого представления, нахождение канонического представителя сводится к нахождению центра масс выпуклого многогранника, и точный алгоритм для этого использует триангуляцию [6], приближенный же — метод Монте-Карло [7].

В рамках теории АБС одной из основных операций, рассматриваемых в отношении фрагментов знаний ФЗ и АБС, являются *апостериорный вывод*. Эта процедура позволяет моделировать интеграцию новой информации о предметной области в АБС/ФЗ. Новая информация формализуется в общем случае также в виде ФЗ и называется *свидетельством*. Для реализации указанных алгоритмов в условиях интервальных оценок требуется решение задач линейного программирования, в то время как в случае скалярных оценок можно ограничиться более простыми матричными операциями. Следовательно, временные затраты алгоритмов, использующих интервальные оценки, могут значительно превышать затраты аналогичных алгоритмов, работающих со скалярными оценками, причем разница может составлять несколько порядков [8]. Данный аспект работы алгоритмов может стать критически важным в условиях ограниченных временных ресурсов или недостаточных вычислительных мощностей, что делает использование объектов со скалярными оценками, а значит и канонического представителя, более предпочтительным в этом случае. Текущая же работа будет направлена, в том числе, на сравнение работы апостериорного вывода с использованием приближенного и точного алгоритмов построения канонического представителя.

III. МЕТОД ПЕРЕХОДА К СКАЛЯРНЫМ ОЦЕНКАМ

В данной главе описывается методология перехода от ФЗ, представленных в виде конъюнктов с интервальными оценками к их представлению в скалярном формате.

На практике оценки чаще всего представляют в виде конъюнктов, однако алгоритмы построения канонического представителя используют представление в виде квантов. Поэтому перед нахождением канонического представителя проводится конвертация ФЗ в формат, использующий кванты [6–7]. После же перехода от интервальных оценок к скалярным идет конвертация представления в формат с конъюнктами. Таким образом, алгоритм нахождения канонического представителя можно представить в виде блок-схемы, указанной на рис. 1.

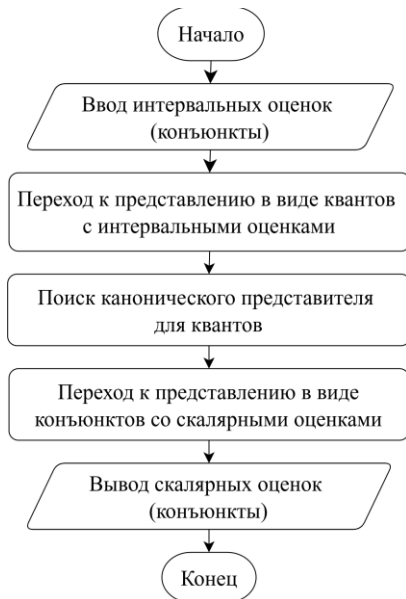


Рис. 1. Блок-схема построения канонического представителя

IV. СРАВНЕНИЕ ТОЧНОГО И ПРИБЛИЖЕННОГО МЕТОДА

Далее будет проведен сравнительный анализ точного и приближенного методов, используемых для нахождения канонического представителя. Поскольку выбор метода влияет на точность получаемых результатов, важно определить, насколько приближенный метод может эффективно заменить точный, особенно в условиях реального применения. Исследование направлено на выявление погрешностей, возникающих при использовании приближенного метода. Для реализации алгоритмов использовался язык программирования *Python*. Оценка погрешностей проходила следующим образом:

- По алгоритму [10] взято 160, 260 и 45 ФЗ, состоящих из 1, 2 и 3 атомов с интервальными оценками соответственно. Взятие различного числа ФЗ обусловлено различной скоростью схождения приближенного алгоритма к точному при увеличении числа ФЗ, что будет видно на рис. 2–4 далее.
- Затем для каждого ФЗ был проведен апостериорный вывод с использованием точного [6] и приближенного [7] алгоритмов для перехода к скалярным оценкам. Свидетельством выступал единый случайный ФЗ, состоящий из 1 атома.
- Погрешность для одного ФЗ оценивалась следующим образом:

$$KP_{Error} = \frac{1}{n} \sum_{i=1}^n |exact_i - approx_i|,$$

где n — число оценок вероятности истинности в ФЗ, $exact_i$ — оценка, найденная точным алгоритмом, $approx_i$ — приближенным.

- Значения погрешности KP_{Error} постепенно усреднялись для возрастающего от 1 до максимального взятого количества ФЗ и указывались на графиках, приведенных на рис. 2–4.

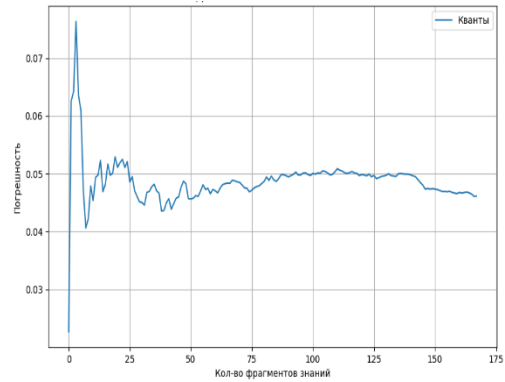


Рис. 2. Погрешность для ФЗ, состоящих из одного атома

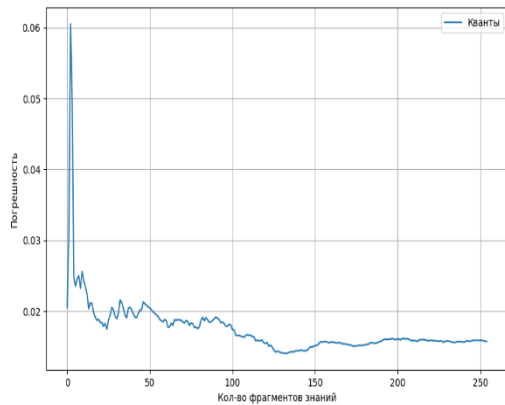


Рис. 3. Погрешность для ФЗ, состоящих из 2 атомов

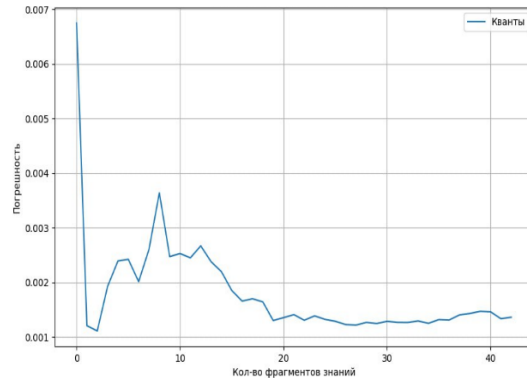


Рис. 4. Погрешность для ФЗ, состоящих из 3 атомов

Как можно видеть из графиков, относительная стабилизация погрешности проходила по разному на различном количестве ФЗ для ФЗ разных размеров. Для ФЗ с 1 и 2 атомами такая стабилизация произошла при 150–200 ФЗ, для ФЗ с 3 атомами — с 30–40 ФЗ. Помимо этого, видно, что стабилизированная погрешность для ФЗ различного размера разная: у ФЗ с 1 атомом она равна примерно 0.0475; у ФЗ с 2-мя атомами — 0.01, с 3-мя — 0.0015. В среднем такая погрешность составила около 0.02 единицы оценки вероятности. Видно, что с увеличением размера ФЗ погрешность уменьшается, однако анализ факторов, определяющих размер погрешности, а также различную скорость стабилизации выходит за рамки статьи и может быть рассмотрен в следующих исследованиях.

V. ПРИМЕР

В качестве примера можно рассмотреть следующие интервальные оценки для фрагмента знаний, состоящего из 2 атомов:

$$p(x_0) = [0.2; 0.4]$$

$$p(x_1) = [0.3; 0.6]$$

$$p(x_0x_1) = [0.25; 0.6]$$

В результате выполнения поиска канонического представителя были получены следующие оценки:

$$p(x_0) = 0.3968$$

$$p(x_1) = 0.4903$$

$$p(x_0x_1) = 0.3241$$

Как можно заметить, полученные скалярные оценки попадают в интервальные оценки, что говорит о том, что полученный ФЗ со скалярными оценками входит в семейство ФЗ, задаваемых ФЗ с интервальными оценками.

VI. ЗАКЛЮЧЕНИЕ

В данной работе описывается алгоритм автоматизации построения канонического представителя ФЗ, а также оценивается погрешность при использовании приближенного метода построения канонического представителя в апостериорном выводе. Согласно проведенному эксперименту, такая оценка составила около 0.02 единицы оценки вероятности. Помимо этого результаты согласуются с утверждением о том, что при увеличении размера ФЗ погрешность при использовании приближенного алгоритма уменьшается.

Дальнейшие исследования могут быть направлены на подробное изучение причин уменьшения погрешности, оптимизации приближенного алгоритма. Результаты данной работы способствуют расширению области применения структур АБС и могут быть полезны, например, в сфере оценки защищенности систем от социоинженерных атак [11–12].

СПИСОК ЛИТЕРАТУРЫ

- [1] Haasler I., Singh R., Zhang Q., Karlsson J., Chen Y. Multi-marginal optimal transport and probabilistic graphical models // IEEE Transactions on Information Theory. 2021. Vol. 67, № 7. P. 4647–4668.
- [2] Yu D., Yang B., Wei Q., Li A., Pan S. A probabilistic graphical model based on neural-symbolic reasoning for visual relationship detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. P. 10609–10618.
- [3] Moolchandani S. Advancing Credit Risk Management: Embracing Probabilistic graphical models in banking // International Journal of Science and Research (IJSR). 2024. Vol. 13, № 6. P. 74–80.
- [4] Forough J., Momtazi S. Sequential credit card fraud detection: A joint deep neural network and probabilistic graphical model approach // Expert Systems. 2022. Vol. 39, № 1. p. e12795.
- [5] Li C., Li Y., Wang C., Dong S., Gao H., Zhao Q., Wu W. The multimedia recommendation algorithm based on probability graphical model // Multimedia tools and applications. 2022. Vol. 81, № 14. P.19035–19050.
- [6] Vyatkin A.A., Abramov M.V. Algebraic bayesian networks: the exact generation of the knowledge pattern canonical representation. // Proceedings of 2024 XXVII International Conference on Soft Computing and Measurements. SCM 2024. 2024. P. 41–45.
- [7] Kharitonov N.A., Tulupyev A.L. Algebraic Bayesian Networks: The Generation of the Knowledge Pattern Canonical Representation // Proceedings of 2021 24-th International Conference on Soft Computing and Measurements. SCM 2021. 2021. P. 144–146.
- [8] Kharitonov N., Vyatkin A., Tulupyev A. Algebraic Bayesian Networks: The Generation of the Network Canonical Representation // Proceedings of the Seventh International Scientific Conference “Intelligent Information Technologies for Industry” (IITI'23). IITI 2023. Lecture Notes in Networks and Systems. 2023. Vol. 777. P. 13–22. doi: 10.1007/978-3-031-43792-2_2.
- [9] Тулупьев А.Л. Алгебраические байесовские сети: локальный логико-вероятностный вывод: Учеб. пособие. СПб.: СПбГУ; ООО Издательство «Анатолия», 2007. 80 с.
- [10] Vyatkin A.A., Tulupyev A.L. Statistical Comparison of the Running Times of Global Posteriori Inference Algorithms in Algebraic Bayesian Networks // 2023 XXVI International Conference on Soft Computing and Measurements, SCM 2023. St. Petersburg, Russia. 2023. P. 24–28.
- [11] 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 2020. St. Petersburg, Russia. 2020. pp. 41–43.
- [12] Khlobystova A.O., Abramov M.V. The models separation of accessrights 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.