Задача распространения виртуального свидетельства в алгебраической байесовской сети

А. В. Шляк[1], А. А. Золотин[2]
Санкт-Петербургский государственный университет,
Санкт-Петербургский институт информатики
и автоматизации РАН
[1]dieseltsame@gmail.com, [2]andrey.zolotin@gmail.com

А. Л. Тулупьев
Санкт-Петербургский институт информатики
и автоматизации РАН
alexander.tulupyev@gmail.com

Аннотация. **Алгебраические байесовские сети представляют собой вероятностную графическую модель, реализующую один из подходов к обработке знаний с неопределенностью. Логико-вероятностный вывод является важной составляющей теории алгебраических байесовских сетей, включая в себя алгоритмы локального и глобального вывода. В данной работе рассматривается глобальный вывод, а именно передача и распространение информации между двумя фрагментами знаний сети. В работе приводится краткий обзор существующих решений и программных реализаций алгоритмов глобального апостериорного вывода, а также формулируются недостатки существующих подходов. Авторы ставят задачу распространения свидетельства и формулируют матрично-векторную форму уравнения, описывающего вычисление апостериорных вероятностей фрагментов знаний после передачи виртуального свидетельства. Работа завершается примером вычисления апостериорных вероятностей и описанием компоненты прототипа комплекса программ, реализующего описанные алгоритмы.**

Ключевые слова: алгебраические байесовские сети; логико-вероятностный вывод; виртуальное свидетельство; идеал конъюнктов; глобальный апостериорный вывод; вероятностные графические модели; знания с неопределенностью

I. Введение

Современные технологии и алгоритмы в области информатики и анализа данных позволяют исследователям за короткий промежуток времени собирать большие объемы информации. Однако, после сбора данных встает нетривиальная задача их обработки, осложняемая зачастую, пусть и неявно, присутствующей в данных неопределенностью. В контексте теории вероятностей источником неопределенности может быть эксперт, дающий оценку некоторого суждения (например, характеризуя некоторое событие как «возможное»), пробелы в данных, заполняемые на основании и в согласованности с имеющимися оценками вероятностей и, наконец, сам подход к оценке вероятности составной пропозициональной формулы. Обработка неопределенности осложняется большими объемами баз знаний, представляемых алгоритмам монолитным объектом, что ведет к экспоненциальному росту объемов вычислений.

Целью данной работы является изучение и развитие матрично-векторного подхода в задаче распространения свидетельства в между фрагментами знаний АБС. Для достижения поставленной цели решается ряд задач: провести обзор существующих решений и выявить присутствующие в них недостатки, предложить подход, описывающий формирование и распространение виртуального свидетельства на матрично-векторном языке, реализовать указанный подход в компоненте прототипа комплекса программ на языке C#.

Алгебраические байесовские сети были предложены профессором В.И. Городецким [1] по результатам многолетних исследований в области знаний с неопределенностью и занимают свое место среди вероятностных графических моделей. Отличительной особенностью АБС является декомпозиция данных на фрагменты знаний (ФЗ), на пересечении которых лежат общие утверждения. В классическом представлении [2] фрагмент знаний характеризуется идеалом конъюнктов C и вектором вероятностей элементов идеала $\mathbf{P}_c$ или функцией вероятности p, заданной на множестве конъюнктов. Особенно стоит отметить, что АБС подразумевает связность системы фрагментов знаний, так как в противном случае она разбивается на несколько компонент связности и рассматривается как различные АБС. Вторичная структура АБС описывает механизм связей [3] между фрагментами знаний, формируя особый граф, вершинами которого являются фрагменты знаний, а ребрами выступают связи между ними [4].

Одной из родственных моделей АБС являются байесовские сети доверия (БСД), находящие множественные применения в медицине [5], оценке рисков [6], финансовом секторе и иных областях [7]. БСД также предоставляют исследователям аппарат логико-вероятностного вывода, включающий в себя априорный и апостериорный виды вывода [8], основанный во многом на принципе d-разделенности. Данный подход, однако, имеет

Работа выполнена при финансовой поддержке РФФИ, проект №18-01-00626 «Методы представления, синтеза оценок истинности и машинного обучения в алгебраических байесовских сетях и родственных моделях знаний с неопределенностью: логико-вероятностный подход и системы графов», рук. А.Л. Тулупьев.

существенный недостаток: при работе с большими моделями отсутствие декомпозиции области знаний требует серьёзных вычислительных мощностей.

II. Релевантные работы

Глобальный вывод в АБС был рассмотрен коллективом на базе ТиМПИ СПИИРАН, включающем А.Л. Тулупьева, А.В. Сироткина, А.А. Фильченкова и других исследователей. Были сформулированы основные требования к АБС, а также описаны основные шаги алгоритма распространения свидетельства [9], которые можно кратко описать следующей последовательностью действий: поступление свидетельства в первый ФЗ, расчет апостериорных оценок вероятностей элементов первого ФЗ, формирование виртуального свидетельства из элементов первого ФЗ, распространение виртуального свидетельства во второй ФЗ.

Кроме того, было предложено [2] функциональное описание формирования виртуального свидетельства. В общем виде его можно представить следующим образом: $p_{[a]}(C_n) = p_a\left(C_n \middle| \left\langle p_{[a]}(C_{n-1})\right\rangle\right)$, где C_i – i-тый ФЗ в сети, представленный идеалом конъюнктов, p – функция вероятности, определенная на идеале конъюнктов, а элемент [a] в нижнем индексе означает апостериорное значение. Данная запись дает общее представление об алгоритме апостериорного вывода, однако скрывает от читателя его детали, важные при программной реализации алгоритма. Кроме того, такая запись во многом осложняет анализ чувствительности алгоритма и оценку его сложности.

III. Распространение виртуального свидетельства

Одним из применений вторичной структуры является глобальный логико-вероятностный вывод, а именно вторая задача апостериорного вывода – распространение влияния свидетельства [9]. Алгоритм распространения свидетельства по алгебраической байесовской сети подразумевает использование декомпозиции предметной области на отдельные фрагменты знаний. Такой подход позволяет существенно сократить объем вычислений в сравнении с проведением апостериорного вывода с погружением сети в объемлющий фрагмент знаний,

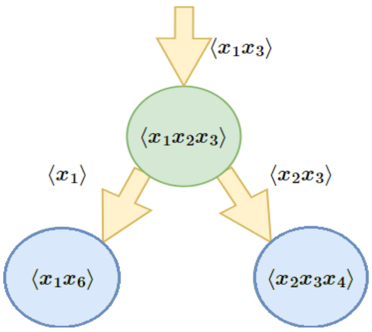

Рис. 1. Распространение свидетельства над алфавитом $\{x_1, x_3\}$ в АБС

дающим экспоненциальный рост количества операций при увеличении мощности алфавита.

Ранее была предложена матрично-векторная форма уравнения для решения задач локального апостериорного вывода [10]. Рассмотрим ациклическую АБС, удовлетворяющую условиям, приведенным выше. Предположим, что в один из фрагментов знаний $\left(C^1, \mathbf{P}_c^1\right)$ поступило стохастическое свидетельство, выраженное фрагментом знаний $\left(C^{ev}, \mathbf{P}_c^{ev}\right)$. Начнем с конца приведенного выше пошагового алгоритма и запишем уравнение передачи указанного свидетельства в следующий ФЗ $\left(C^2, \mathbf{P}_c^2\right)$. Основываясь на результатах, полученных для локального апостериорного вывода, получим следующий результат [11]:

$$\mathbf{P}_c^{2,a} = \sum_{i=0}^{2^{n'}-1} \frac{\mathbf{T}^{\langle \mathrm{GInd}(i,m),\mathrm{GInd}(2^{n'}-1-i,m)\rangle}\mathbf{P}_c^2}{\left(\mathbf{r}^{\langle \mathrm{GInd}(i,m),\mathrm{GInd}(2^{n'}-1-i,m)\rangle}, \mathbf{P}_c^2\right)}\mathbf{I}_n\mathbf{P}_c^{sep}[i], \quad (2)$$

где $\mathbf{P}_c^{sep}$ – вектор вероятностей виртуального свидетельства. Матрица $\mathbf{I}_n$ является матрицей перехода от вектора вероятностей идеала конъюнктов $\mathbf{P}_c$ к вектору вероятностей набора квантов $\mathbf{P}_q : \mathbf{P}_q = \mathbf{I}_n\mathbf{P}_c$. Подробное построение матрицы $\mathbf{T}^{\langle i,j\rangle}$ и вектора-редистрибьютора $\mathbf{r}^{\langle i,j\rangle}$ описано в работе [10]. Вектор $\mathbf{P}_c^{sep}$ получается выделением из вектора $\mathbf{P}_c^1$ вероятностей элементов, принадлежащих обоим ФЗ (принадлежащих сепаратору). Таким образом, вектор $\mathbf{P}_c^{sep}$ можно выразить как вектор $\mathbf{P}_c^1$, умноженный на проекционную матрицу $\mathbf{Q} : \mathbf{P}_c^{sep} = \mathbf{Q}\mathbf{P}_c^1$.

Благодаря предложенному разбиению задачи апостериорного вывода на шаги, нам удалось выделить элемент уравнения (матрицу $\mathbf{Q}$), которого не хватает для полноты описания. Тем не менее, задача формирования матрицы $\mathbf{Q}$ остается по-прежнему открытой.

IV. Программная реализация

Матрично-векторная форма описания уравнений логико-вероятностного вывода упрощает их программную реализацию. В целях автоматизации алгоритмов ЛВВ были разработаны библиотеки для работы с алгебраическими байесовскими сетями на языках Java и C++ [12]. Однако, теория АБС продолжает развиваться, в частности были предложены новые уравнения для решения задач локального ЛВВ, что повлекло создание комплекса программ AlgBNetModeller [13].

Программный комплекс содержит интерфейсы и классы для создания различных моделей ФЗ с интервальными, скалярными или бинарными оценками вероятностей элементов. Структура Inferrer позволяет проверить ФЗ на непротиворечивость и провести локальный априорный ЛВВ. Структура Propagator отвечает за апостериорный логико-вероятностный вывод для различных видов ФЗ и свидетельств. Для решения задач линейного программирования используется библиотека lp_solve55. Также есть структура Alphabet для работы с алфавитом, MatrixTransform, реализующая функциональность матриц перехода $\mathbf{I}_n$ и $\mathbf{J}_n$, и другие вспомогательные классы.

За распространение виртуального свидетельства отвечают классы VirtualEvidenceScalarConjPropagator и VirtualEvidenceIntervalConjPropagator [14], содержащие методы распространения для неточного и стохастического виртуальных свидетельств. Сигнатуры методов представлены на рис. 2. Интерфейсы ScalarConjKP_I и IntervalConjKP_I отвечают за хранение ФЗ со скалярными и интервальными оценками соответственно.

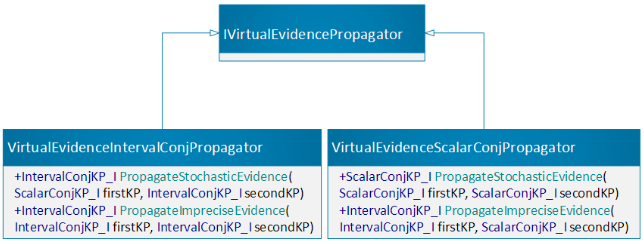

Рис. 2. Диаграмма классов для распространения виртуального свидетельства

Важнейшим шагом распространения виртуального свидетельства из одного фрагмента знаний в другой как раз является его формирование. Из двух фрагментов знаний нужно выделить общие элементы, находящиеся на пересечении. За это отвечает метод расширения GetSeparator для интерфейсов ScalarConjKP_I и IntervalConjKP_I.

Стоит отметить, что при программной реализации алфавит удобнее задавать с помощью глобального индекса – числа в двоичной записи, в котором 1 на i-том месте обозначает, что i-тый элемент алфавита входит в фрагмент знаний. Глобальный индекс сепаратора можно получить с помощью операции побитового И глобальных индексов первого и второго фрагментов знаний. Данную операцию можно провести с помощью следующего фрагмента кода:

```
for (int index = 0; index <= KPGlobalIndex; index++) {
    if ((index & KPGlobalIndex) == index) {
        if ((index & separatorGlobalIndex) == index)  {
            probability[separatorIndex] =
firstKP.GetPointEstimate(probIndex);
            probIndex++;
            separatorIndex++;
        }
        else probIndex++;
    }
}
```

Представленный выше цикл проверяет все возможные числа index, которые могут соответствовать элементам вектора вероятностей, это числа от 0 до KPGlobalIndex – глобального индекса первого ФЗ. Первое условие проверяет это соответствие, второе условие аналогично проверяет, соответствует ли индекс элементу сепаратора. Если условия верны, то значение вероятности сохраняется в массив вероятностей probability. Индексы текущих элементов в векторе вероятностей фрагмента знаний и векторе вероятностей сепаратора задаются с помощью переменных probIndex и separatorIndex соответственно. Стоит отметить, что данный цикл реализует функциональность матрицы перехода $\mathbf{Q}$.

V. Пример распространения виртуального свидетельства

Рассмотрим пример распространения виртуального свидетельства из одного фрагмента знаний в другой. Пусть даны два фрагмента знаний над алфавитами $A_1 = \{x_1, x_2\}$ и $A_2 = \{x_2, x_3\}$, пусть в первый фрагмент знаний поступило стохастическое свидетельство над алфавитом $\{x_1\}$. Пусть ФЗ и свидетельство имеют следующие оценки вероятностей:

$$\mathbf{P}_c^1 = \begin{pmatrix} 1 \\ p(x_1) \\ p(x_2) \\ p(x_2x_1) \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \\ 0.7 \\ 0.3 \end{pmatrix}; \mathbf{P}_c^2 = \begin{pmatrix} 1 \\ p(x_2) \\ p(x_3) \\ p(x_3x_2) \end{pmatrix} = \begin{pmatrix} 1 \\ 0.3 \\ 0.7 \\ 0.1 \end{pmatrix};$$

$$\mathbf{P}_c^{ev} = \begin{pmatrix} 1 \\ p(x_1) \end{pmatrix} = \begin{pmatrix} 1 \\ 0.4 \end{pmatrix}.$$

После распространения свидетельства оценки вероятностей элементов первого фрагмента знаний будут следующими: $\mathbf{P}_c^{1,a} = \begin{pmatrix} 1 & 0.4 & 0.72 & 0.24 \end{pmatrix}^{\mathrm{T}}$.

Виртуальное свидетельство будет построено над алфавитом $A_{sep} = A_1 \cap A_2 = \{x_1, x_2\} \cap \{x_2, x_3\} = \{x_2\}$ и будет содержать вектор оценок вероятностей элементов $\mathbf{P}_c^{sep}$, после его распространения с помощью уравнения (2) вектор вероятностей второго ФЗ будет равен $\mathbf{P}_c^{2,a}$:

$$\mathbf{P}_c^{sep} = \begin{pmatrix} 1 \\ p(x_2) \end{pmatrix} = \begin{pmatrix} 1 \\ 0.72 \end{pmatrix}, \mathbf{P}_c^{2,a} = \begin{pmatrix} 1 & 0.72 & 0.48 & 0.24 \end{pmatrix}^{\mathrm{T}}.$$

Данный результат распространения свидетельства можно получить с помощью следующего фрагмента кода:

```
// Creating KP, evidence and propagators
var firstKP = new ScalarConjKP(3, new[] { 1, 0.5, 0.7,
0.3 });
var secondKP = new ScalarConjKP(6, new[] { 1, 0.3, 0.7,
0.1 });
var evidence = new ScalarConjKP(1, new[] { 1, 0.4 });
var localPropagator = new
StochasticScalarConjunctsLocalPropagator();
var virtualEvPropagator = new
VirtualEvidenceScalarConjPropagator();
// Propagating stochasic evidence in fisrt KP
localPropagator.setPattern(firstKP);
localPropagator.propagate(evidence);
```

```
var firstKPwithAposterioriEst =
localPropagator.getResult();
// Propagating virtual evidence in second KP
var secondKPwithAposterioriEst = virtualEvPropagator
.PropagateStochasticEvidence(firstKPwithAposterioriEst,
secondKP);
```

Вывод на консоль массива оценок вероятностей firstKPwithAposterioriEst.GetPointEstimate() даст следующий результат: 1, 0,4 0,72 0,24; secondKPwithAposterioriEst.GetPointEstimate() выведет следующий массив: 1, 0,72, 0,48 0,24.

VI. Заключение

Приведенные в работы алгоритмы глобального вывода во многом основываются на полученных ранее и более глубоко исследованных матрично-векторных уравнениях локального апостериорного вывода. Сформулированное уравнение передачи свидетельства из одного ФЗ в другой вводит матрицу $\mathbf{Q}$, которую предстоит исследовать в последующих работах по данной тематике. Открытым вопросом по прежнему остается сам алгоритм формирования матрицы $\mathbf{Q}$. Предпочтительно реализовать вычисление матрицы с помощью произведения матриц меньших размерностей, так как это позволит в дальнейшем использовать соответствующие структуры при имплементации алгоритмов в комплексе программ.

Реализация алгоритмов основывается на описанных в работе теоретических результатах. Предложенная структура классов является модулем объемлющей математической библиотеки, реализующей алгоритмы логико-вероятностного вывода в АБС и предоставляющей доступ к публичному контракту [13]. Благодаря публичному контракту, математическая библиотека может быть переиспользована при разработке веб-приложения, направленного на визуализацию структур АБС, их синтез и проведение вывода в над ними [15].

Полученные результаты являются первым шагом в сторону строгого математического изложения аппарата глобального вывода. Такое изложение позволит исследовать и дать оценку чувствительности и устойчивости модели, что оказывается важным параметром не только с теоретической точки зрения, подтверждая корректность модели, но и с практической, давая возможность оценить претенциозность к собираемым данным. Иным направлением дальнейших исследований являются АБС над альтернативными моделями фрагментов знаний – идеалом дизъюнктов и множеством пропозиций-квантов. Применение данных моделей обусловлено отличиями в структуре данных, над которыми строится АБС. Кроме того, в уравнениях как локального так и глобального вывода перед исследователями стоит задача строгого математического описания процесса, выполняемого функцией GInd(), являющейся по своей сути проекцией одного множества на другое.

Отметим также, что АБС могут найти применение и в рамках исследования, проектирования и разработки методологии защиты пользователей системы от социоинженерных атак. Большой объем информации и множество неизвестных (или частично известных) параметров создают благоприятные условия для применения алгоритмов синтеза структуры и алгоритмов логико-вероятностного вывода, описанных в теории АБС.

Список литературы

[1] Городецкий В.И. Алгебраические байесовские сети—новая парадигма экспертных систем //Юбилейный сборник трудов институтов Отделения информатики, вычислительной техники и автоматизации РАН. 1993. Т. 2. С. 120-141.

[2] Тулупьев А.Л., Сироткин А.В., Николенко С.И. Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. Изд-во С.-Петерб. ун-та, СПб., 2009. 400 С.

[3] Тулупьев А.Л. Дерево смежности с идеалами конъюнктов как ациклическая алгебраическая байесовская сеть //Труды СПИИРАН. 2006. Т. 1. №. 3. С. 198-227.

[4] Фильченков А.А., Тулупьев А.Л., Сироткин А.В. Особенности анализа вторичной структуры алгебраической байесовской сети //Труды СПИИРАН. 2010. Т. 1. №. 12. С. 97-118.

[5] Lucas P. J. F., Van der Gaag L.C., Abu-Hanna A. Bayesian networks in biomedicine and health-care. // Artificial Intelligence in Medicine. V.30. 2004. P. 201–214.

[6] Wang J.F., Liu X.C., Liao Y., Xu B. A. A new method for assessing the risk of infectious disease outbreak // Scientific reports. 2017 V. 7. P. 40084.

[7] Liu F., Song Y.Z., Xiang T., Hospedales T., Yu Q., Yang Y.X.M. A deep neural network that beats humans // International Journal of Computer Vision. 2016. P. 1–15.

[8] Sucar L.E. Bayesian Networks: Representation and Inference //Probabilistic Graphical Models. Springer, London, 2015. C. 101-136.

[9] Тулупьев А.Л. Алгебраические байесовские сети: глобальный логико-вероятностный вывод в деревьях смежности: Учеб. пособие. Элементы мягких вычислений. СПб.: СПбГУ; ООО Издательство «Анатолия», 2007. 40 с.

[10] Золотин, А.А., Тулупьев А.Л., Сироткин А.В. Матрично-векторные алгоритмы нормировки для локального апостериорного вывода в алгебраических байесовских сетях // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15, № 1. С. 78–85.

[11] Золотин А.А., Шляк А.В., Тулупьев А.Л. Пропагация виртуального стохастического свидетельства в алгебраических байесовских сетях: алгоритмы и уравнения // Труды VII Всероссийской Научно-практической Конференции Нечеткие Системы, Мягкие Вычисления и Интеллектуальные Технологии (НСМВИТ-2017). Т. 2. С. 96–106.

[12] Тулупьев А.Л. Алгебраические байесовские сети: реализация логико-вероятностного вывода в комплексе java-программ //Труды СПИИРАН. 2009. №. 8. С. 191–232.

[13] Мальчевская Е.А., Золотин А.А. Логико-вероятностный вывод в АБС: архитектура и примеры использования программного комплекса на языке C# // Гибридные и Синергетические Интеллектуальные Системы. Материалы III Всероссийской Поспеловской конференции с международным участием /Под редакцией А.В. Колесникова. 2016. С. 181–187.

[14] Пат. РФ № 2017664086. / А.В. Шляк, А.А. Золотин, А.Л. Тулупьев. Algebraic Bayesian Network Virtual Evidence Propagators, Version 01 for CSharp (AlgBN VE Propagators cs.v.01); Опубл. 21.02.2018 Бюл. №3.

[15] Золотин А.А., Левенец Д.Г., Мальчевская Е.А., Зотов М.А., Бирилло А.И., Березин А.И., Иванова А.В., Тулупьев А.Л. Алгоритмы обработки и визуализации алгебраических байесовских сетей. Образовательные технологии и общество. 2017. Т. 1. С. 446–457.