5 КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004.8

ПОВЫШЕНИЕ БЫСТРОДЕЙСТВИЯ АЛГОРИТМА ОЦЕНКИ НАБЛЮДАЕМОЙ ПОСЛЕДОВАТЕЛЬНОСТИ В СКРЫТЫХ МАРКОВСКИХ МОДЕЛЯХ НА ОСНОВЕ АЛГЕБРАИЧЕСКИХ БАЙЕСОВСКИХ СЕТЕЙ

М.Я. Пинский, А.В. Сироткин, А.Л. Тулупьев, А.А. Фильченков

Скрытые марковские модели и алгебраические байесовские сети представляют собой вероятностные графические модели, а потому во многом похожи. Скрытые марковские модели получили широкое применение, в то время как алгебраические байесовские сети пока не столь распространены, однако их аппарат позволяет моделировать и решать задачи скрытых марковских моделей. Рассмотрен вопрос ускорения решения первой задачи скрытых марковских моделей на основе методов, применяющихся в алгебраических байесовских сетях. Предложен алгоритм для оценки вероятности наблюдаемой последовательности в бинарных линейных по структуре скрытых марковских моделях с помощью апостериорного вывода алгебраической байесовской сети.

Ключевые слова: скрытые марковские модели, алгебраические байесовские сети, бинарные линейные по структуре скрытые марковские модели, вероятностные графические модели.

Введение

Вероятностные графические модели – скрытые марковские модели (СММ) и алгебраические байесовские сети (АБС) – используются для изучения различных процессов в таких областях, как распознавание речи, теория информации, машинный перевод, молекулярная биология [1–6]. СММ – более развитый и широко известный инструмент для моделирования временны́х рядов. Их используют во многих современных системах распознавания речи [4], в большинстве приложений вычислительной молекулярной биологии [7], в сжатии информации [8], в системах статистического машинного перевода [5], приложениях компьютерного зрения [9].

АБС – это одна из математических моделей баз фрагментов знаний с неопределенностью. Она формализует знания (с неопределенностью) при помощи вероятностной логики. Развитие аппарата АБС осуществлялось с 1980-х г.г., и на сегодняшний день в теории АБС существуют алгоритмы для решения различных задач, однако АБС все еще редко используются для практических целей [10]. На текущий момент АБС обладают развитым аппаратом логико-вероятностного вывода [11–15] и набором средств автоматического обучения, который находится на стадии развития [16–18].

Области применения АБС и СММ схожи, поэтому возникает вопрос о возможности представления одной модели через другую. Это может быть полезно для использования разработок, полученных в одной из них, для более широкого круга задач.

Цель работы – ускорение известного алгоритма [19] решения первой задачи для СММ с помощью апостериорного вывода АБС.

Скрытые марковские модели

Определения, связанные со СММ, будут вводиться по [2, 4, 6]. СММ – модель, состоящая из следующих объектов:

1. набор возможных значений скрытых состояний $S = \{s_1, s_2, \dots s_N\}$;
2. последовательность скрытых состояний во времени $Q = \{q_1, q_2, \dots q_T\}$;
3. матрица переходных вероятностей $A = \{a_{ij}\}$, $a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$, $1 \le i, j \le N$;
4. вектор начального распределения $\pi = \{\pi_i\}$, $\pi_i = P(q_1 = s_i)$, $1 \le i \le N$;
5. алфавит возможных значений наблюдений $V = \{v_1, v_2, \dots v_M\}$;
6. последовательность наблюдений во времени $O = \{o_1, o_2, \dots o_T\}$;
7. матрица вероятностей наблюдений $B = \{b_j(k)\}$, $b_j(k) = P(v_k = o_t | s_j = q_t)$, $1 \le j \le N$, $1 \le k \le M$,

и обладающая следующими свойствами:

1. $P(q_{t+1} | q_t, q_{t-1}, q_{t-2}, \dots, q_1) = P(q_{t+1} | q_t)$ – марковское свойство;
2. $P(o_t | o_1, \dots o_T, q_1, \dots, q_T) = P(o_t | q_t)$ – зависимость текущего наблюдения только от текущего состояния,

где $O = \{o_1, o_2, \dots o_T\}$, $Q = \{q_1, q_2, \dots q_T\}$ – последовательности наблюдений и состояний соответственно.

В расчетах, связанных со СММ, пользуются представлением СММ в виде набора матриц вероятностей: $\mu = (A, B, \pi)$.

В теории СММ сформулированы три основных задачи.

1. *Правдоподобие наблюдений.* Дана СММ с известными матрицами вероятностей. Определить вероятность поступающей последовательности наблюдений во времени относительно этой СММ.
   Формальная постановка задачи: дана последовательность наблюдений $O = \{o_1, o_2, \ldots o_T\}$ и модель $\mu = (A, B, \pi)$. $P(O|\mu) = ?$ Для данной задачи существуют различные решения, например алгоритм «вперед-назад» [19].
2. *Декодирование скрытой последовательности.* Даны СММ с оценками вероятности и поступившая последовательность наблюдений. Требуется определить наиболее вероятную последовательность скрытых состояний. Формальная формулировка задачи: дана последовательность наблюдений $O = \{o_1, o_2, \ldots o_T\}$ и модель $\mu = (A, B, \pi)$. Найти наиболее вероятную последовательность скрытых состояний $Q = \{q_1, q_2, \ldots q_T\}$, соответствующую данной последовательности наблюдений. Данная задача решается с помощью алгоритма Витерби [20].
3. *Обучение СММ.* Изменить (настроить) матрицы вероятностей СММ таким образом, чтобы максимизировать вероятности поступающего набора последовательностей наблюдений. Иначе говоря, требуется обучить СММ на наборе тренировочных последовательностей наблюдений. Формальная формулировка задачи: настроить параметры модели $\mu = (A, B, \pi)$ так, чтобы максимизировать $P(O|\mu)$. Данную задачу можно решать с помощью алгоритма Баума–Вэлха [21].

Для преобразования в АБС в работе будем использовать бинарные линейные по структуре СММ. Это такие СММ, у которых могут быть только два скрытых состояния $(S = \{s_1, s_2\})$ и два вида наблюдений $(V = \{v_1, v_2\})$. В последнем будет иметь место $S = V = \{0,1\} = \{true, false\}$.

Алгебраические байесовские сети

Определения, связанные с теорией АБС, будут вводиться по [10, 22]. АБС – логико-вероятностная графическая модель баз фрагментов знаний с неопределенностью [10]. Фрагмент знаний представляется в виде идеала конъюнктов с оценками их истинности.

Пусть T – конечный набор элементарных пропозиций; $S = \{s_1, s_2, \ldots, s_k\}$ – непустое подмножество T; $S^{\Diamond} = \{v_1 v_2 \ldots v_r : v_1, v_2 \ldots v_r \in S, r = 0 \ldots k\}$ – конъюнкты (множество положительно означенных конъюнкций над S); $S^{\Delta} = S^{\Diamond} \setminus \{e\}$ – идеал конъюнкта (множество положительно означенных конъюнктов без пустого конъюнкта). В идеале существует один максимальный элемент (максимальная по длине конъюнкция) и множество минимальных элементов (одноатомных конъюнкций).

Квант $Q = \{\tilde{x}_0 \tilde{x}_1 \ldots \tilde{x}_{n-1}\}$ – конъюнкция, которая для любой атомарной переменной из алфавита содержит либо ее формулу, либо отрицание.

Теперь введем нумерацию на конъюнктах и квантах. Каждому конъюнкту из идеала $\{x_{i_1} x_{i_2} \ldots x_{i_k} | 0 \le i_1 < \ldots < i_k \le n-1, k \le n\}$ можно сопоставить номер вида $2^{i_1} + 2^{i_2} + \ldots + 2^{i_k}$. Обозначим через c_i конъюнкт с порядковым номером i. Выделим из кванта положительную часть. Номер получившегося конъюнкта будет номером кванта. Обозначим через q_i квант с порядковым номером i.

Далее введем вероятность на конъюнктах и квантах. Вероятности, относящиеся к фрагменту знаний, удобно упорядочивать по номерам конъюнктов и квантов и представлять в виде векторов. Выделяют две структуры алгебраических байесовских сетей – первичную и вторичную. Первичная структура АБС – это база фрагментов знаний (ФЗ). Вторичная структура АБС – это связи между ФЗ. Вторичная структура АБС определяется набором вершин и набором ребер между ними. Вершины – это фрагменты знаний, они однозначно задаются глобальным индексом минимального конъюнкта. Ребра однозначно задаются двумя вершинами.

Представление бинарных линейных по структуре СММ в виде АБС

Представление вводится по [23], где приведено формальное доказательство корректности сведения бинарных линейных по структуре СММ к АБС. Отметим, что любая СММ может быть приведена к АБС серией аналогичных преобразований. Рассмотрим простейшую линейную бинарную СММ в четыре момента времени (рис. 1). Матрицы данной модели $\mu = (A, B, \pi)$ будут иметь следующий вид:

$$A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} = \begin{pmatrix} p(x_i|x_i) & p(x_i|\bar{x}_i) \\ p(\bar{x}_i|x_i) & p(\bar{x}_i|\bar{x}_i) \end{pmatrix}, \quad B = \begin{pmatrix} b_0(v_0) & b_0(v_1) \\ b_1(v_0) & b_1(v_1) \end{pmatrix} = \begin{pmatrix} p(o_i|x_i) & p(\bar{o}_i|x_i) \\ p(o_i|\bar{x}_i) & p(\bar{o}_i|\bar{x}_i) \end{pmatrix}, \quad \pi = \begin{pmatrix} \pi_0 \\ \pi_1 \end{pmatrix} = \begin{pmatrix} p(x_i) \\ p(\bar{x}_i) \end{pmatrix}.$$

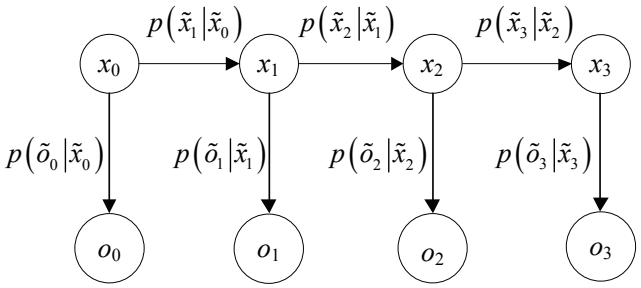

Рис. 1. Линейная бинарная СММ в четыре момента времени

Теперь построим АБС, соответствующую рассмотренной СММ (рис. 2).

Для нумерации вершин необходимо ввести алфавит. В алфавите соответствующей АБС будем чередовать o_i и x_i, начиная с $i = 0$: $\{o_0, x_0, o_1, x_1, \ldots o_{N-1}, x_{N-1}\}$.

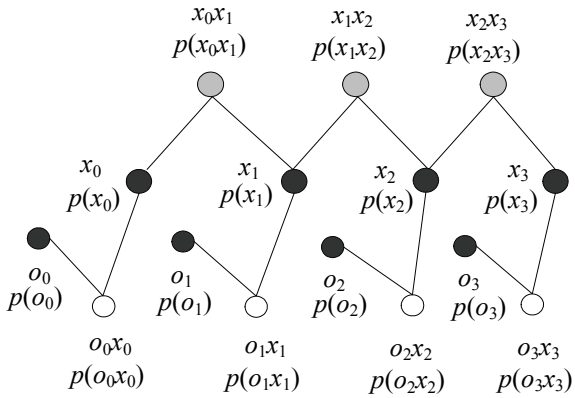

Рис. 2. АБС, соответствующая рассмотренной СММ. (Серым отмечены узлы, соответствующие одноатомным конъюнктам, белым – соответствующие двухатомным конъюнктам)

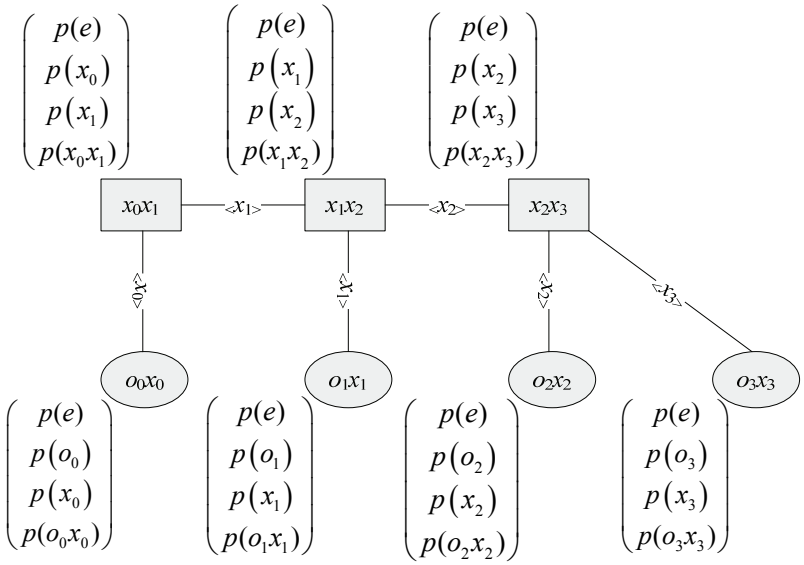

Рис. 3. Соответствующая АБС, изображенная как база ФЗ; изображены узлы и векторы вероятностей в них

Данная АБС будет содержать ФЗ вида $\left\{\left\{o_i, x_i\right\}^{\Delta}, \left\{x_i, x_{i+1}\right\}^{\Delta}_{i=0}{}^{N-2}, \left\{o_{N-1}, x_{N-1}\right\}^{\Delta}\right\}$ (рис. 3).

Решение первой задачи для СММ через АБС

В ходе исследований удалось установить, что первая задача для СММ эквивалентна первой задаче апостериорного вывода для АБС [4].

Первая задача СММ**:** Дана последовательность наблюдений $O = \{o_1, o_2, ...o_T\}$ и модель $\mu = (A, B, \pi)$. Какова вероятность наблюдаемой последовательности при условии данной модели $P(O \mid \mu) = ?$

В терминах АБС данная задача будет формулироваться следующим образом. Поступает детерминированное свидетельство, например $e = \tilde{o}_0\tilde{o}_1\tilde{o}_2...\tilde{o}_T$, каким-то образом означенное. Требуется оценить вероятность данного свидетельства $P(e) = ?$

Рассмотрим какое-либо конкретно-означенное детерминированное свидетельство $e = \tilde{o}_0\tilde{o}_1\tilde{o}_2...\tilde{o}_T$. В теории АБС стандартный алгоритм первой задачи апостериорного вывода может пропагировать (распространять влияние) только свидетельство, полностью лежащее в ФЗ. Данное принадлежит $T+1$ фрагменту знаний. Совершим преобразование, воспользовавшись правилом Байеса:

$$P(\tilde{o}_0\tilde{o}_1\tilde{o}_2...\tilde{o}_T) = P(\tilde{o}_0\tilde{o}_1\tilde{o}_2...\tilde{o}_{T-1} \mid \tilde{o}_T) \cdot P(\tilde{o}_T) = ...$$

$$= P(\tilde{o}_0 \mid \tilde{o}_1\tilde{o}_2...\tilde{o}_T) \cdot P(\tilde{o}_1 \mid \tilde{o}_2...\tilde{o}_T) \cdot ... \cdot P(\tilde{o}_{T-1} \mid \tilde{o}_T) \cdot P(\tilde{o}_T).$$

Теперь исходная вероятность состоит из произведения вероятностей свидетельств, полностью лежащих в соответствующих ФЗ. Будем пропагировать, начиная с крайнего правого подсвидетельства ($\tilde{o}_T$). Оно будет поступать на вход к ФЗ $\tilde{o}_T x_T$. После его пропагации на соседний ФЗ, он будет иметь новые апостериорные оценки вероятностей $P^{\tilde{o}_T}(...)$. Для этих оценок вероятностей будем пропагировать $\tilde{o}_{T-1}$, поступающее в ФЗ $\tilde{o}_{T-1}x_{T-1}$, которое снова изменит оценки вероятностей для следующего ФЗ на апостериорные $P^{\tilde{o}_T\tilde{o}_{T-1}}(...)$, и т.д. Каждый раз будем пропагировать на соседний ФЗ единичные подсвидетельства, двигаясь справа налево. В конце останется только перемножить вероятности единичных подсвидетельств. Данный алгоритм эквивалентен описанному в [1], так как свидетельство достаточно пропагировать только на следующий ФЗ. Его апостериорная оценка изменит оценки следующих свидетельств так же, как если бы свидетельство пропагировали на всю сеть, потому что для любого ФЗ $\tilde{o}_i x_i$ справедливо соотношение

$$P(\tilde{o}_i \mid \tilde{o}_{i+1}...\tilde{o}_T) = P(\tilde{o}_i \mid \tilde{o}_{i+1}) P(\tilde{o}_{i+1} \mid \tilde{o}_{i+2}...\tilde{o}_T).$$

Заключение

Известно, что СММ могут быть представлены как частный случай динамических байесовских сетей доверия, которые, в свою очередь, могут быть преобразованы в АБС. Отсюда возник естественный вопрос о прямой связи между СММ и АБС, который и был частично изучен в данной работе**.**

В работе приведены теория СММ, в том числе первая задача СММ, теория АБС, в частности, апостериорный вывод АБС, а также представление бинарной линейной по структуре СММ в виде АБС. Доказана эквивалентность апостериорного вывода для АБС и первой задачи СММ. Дан улучшенный (ускоренный) по сравнению с [19] алгоритм решения первой задачи СММ, состоящей в оценке вероятности последовательности наблюдений, в терминах апостериорного вывода АБС. Сложность была уменьшена в n раз, где n – количество состояний в СММ. Таким образом, приведенный алгоритм решения работает за полиномиальное от длины входа время. Точную оценку установить не представляется возможным, так как вопрос сложности пропагации свидетельств на данный момент недостаточно изучен.

Первая задача была решена через АБС, что является примером использования АБС в теории СММ и может упростить дальнейшее развитие теории АБС в применении к СММ. Однако для этого необходимо проведение исследований. За рамками работы остаются вопросы, связанные со второй и третьей задачами СММ, которые также разрешимы в теории АБС, однако еще не было создано конкретных алгоритмов решения.

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. Николенко С.И., Тулупьев А.Л. Самообучающиеся системы. – М.: МНЦМО, 2009. – 288 с.
2. Cowell R.G., Dawid A.P., Lauritzen S.L., Spiegelhalter D.J. Probabilistic Networks and Expert Systems. – NY: Springer-Verlag, 1999. – 321 p.
3. Huang X., Acero A., Hsiao-Wuen Hon Spoken Language Processing. – Prentice Hall, 2001. – 1008 p.
4. Huang X., Jack M. and Y. Ariki. Hidden Markov Models for Speech Recognition. – Edinburgh University Press, 1990. – 276 p.
5. Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. – 2-d edition. – Prentice-Hall, 2009. – 944 p.

6. Stengel M. Introduction to Graphical Models, Hidden Markov Models and Bayesian Networks. Department of Information and Computer Sciences Toyohashi University of Technology Toyohashi, 441-8580. – Japan, 2003. – 46 p.
7. da-Silva C.Q. Hidden Markov models applied to a subsequence of the Xylella fastidiosa genome // Genet. Mol. Biol. – São Paulo, Dec. 2003. – V. 26. – № 4. – P. 529–535.
8. Li J., Gray R.M. Image Segmentation and Compression Using Hidden Markov Models. – 1-st edition. – Springer, 2000. – 141 p.
9. Bunke H., Caelli T. Hidden Markov Models Applications in Computer Vision. Series in Machine Perception and Artificial Intelligence. – World Scientific, 2001. – V. 45. – 244 p.
10. Тулупьев А.В., Николенко С.И., Сироткин А.В. Байесовские сети: логико-вероятностный подход. – СПб: Наука, 2006. – 608 с.
11. Сироткин А.В. Модели, алгоритмы и вычислительная сложность синтеза согласованных оценок истинности в алгебраических байесовских сетях // Информационно-измерительные и управляющие системы. – 2009. – № 11. – С. 32–37.
12. Тулупьев А.Л. Алгебраические байесовские сети: реализация логико-вероятностного вывода в комплексе java-программ // Труды СПИИРАН. – СПб: Наука, 2009. – Вып. 8. – С. 191–232.
13. Тулупьев А.Л. Алгебраические байесовские сети: система операций глобального логико-вероятностного вывода // Информационно-измерительные и управляющие системы. – 2010. – № 11. – С. 65–72.
14. Тулупьев А.Л. Апостериорные оценки вероятностей в идеале конъюнктов // Вестник СПбГУ. – 2010. – Сер. 10. – Вып. 1. – С. 95–104.
15. Тулупьев А.Л., Сироткин А.В., Николенко С.И. Байесовские сети доверия: логико-вероятностный вывод в ациклических направленных графах. – СПб: Изд-во СПбГУ, 2009. – 400 с.
16. Опарин В.В., Фильченков А.А., Тулупьев А.Л., Сироткин А.В. Матроидное представление семейства графов смежности над набором фрагментов знаний // Научно-технический вестник СПбГУ ИТМО. – 2010. – № 4. – С. 73–76.
17. Фильченков А.А., Тулупьев А.Л., Сироткин А.В. Компаративный анализ клик минимальных графов смежности алгебраических байесовских сетей // Труды СПИИРАН. – СПб: Наука, 2010. – Вып. 2. – С. 87–105.
18. Тулупьев А.Л. Задача локального автоматического обучения в алгебраических байесовских сетях: логико-вероятностный подход // Труды СПИИРАН. – СПб: Наука, 2008. – Вып. 7. – С. 11–25.
19. Момзикова М.П., Великодная О.И., Пинский М.Я., Сироткин А.В., Тулупьев А.Л., Фильченков А.А. Оценка вероятности наблюдаемой последовательности в бинарных линейных по структуре скрытых марковских моделях с помощью апостериорного вывода в алгебраических байесовских сетях // Труды СПИИРАН. – СПб: Наука, 2010. – Вып. 2. – С. 122–142.
20. Forney D.G. The Viterbi Algorithm // Proceedings of the IEEE. – 1973. – V. 61. – № 3. – P. 268–278.
21. Welch L.R. Hidden Markov Models and the Baum-Welch Algorithm // IEEE Information Theory Society Newsletter. – 2003. –V. 53. – № 4. – P. 10–13.
22. Тулупьев А.Л. Алгебраические байесовские сети. Логико-вероятностный подход к моделированию баз знаний с неопределенностью. – СПб: СПИИРАН, 2000. – 292 с.
23. Момзикова М.П., Великодная О.И., Пинский М.Я., Сироткин А.В., Тулупьев А.Л., Фильченков А.А. Представление бинарных линейных по структуре скрытых марковских моделей в виде алгебраических байесовских сетей // Труды СПИИРАН. – СПб: Наука, 2010. – Вып. 1. – С. 134–150.

Пинский Михаил Яковлевич – Санкт-Петербургский государственный университет информационных технологий, механики и оптики, студент, mikhailpinsky@gmail.com

Сироткин Александр Владимирович – Санкт-Петербургский институт информатики и автоматизации РАН, мл. научный сотрудник, alexander.sirotkin@gmail.com

Тулупьев Александр Львович – Санкт-Петербургский институт информатики и автоматизации РАН, доктор физ.-мат. наук, доцент, зав. лабораторией, alexander.tulupyev@gmail.com

Фильченков Андрей Александрович – Санкт-Петербургский институт информатики и автоматизации РАН, мл. научный сотрудник, aaafil@mail.ru