УДК 62-50

А.Н. ТЕРЕНТЬЕВ, П.И. БИДЮК

ЭВРИСТИЧЕСКИЙ МЕТОД ПОСТРОЕНИЯ БАЙЕСОВСКИХ СЕТЕЙ

Abstract: *Bayesian networks are the instrument, which is widely used for classification task when performing data analysis. The network structure is a NP-hard problem. The paper presents a heuristic method for constructing Bayesian network, based on using mutual information between all vertexes and as an estimation function in each iteration to use description of minimum long. For calculation of the error of learning the formula of the structure difference is proposed. Basic definitions and correspondent illustrative examples are given.*
Key words: *Bayesian network, machine learning, minimum description length (MDL) principle, mutual information, heuristic method of learning.*

Анотація: *Байєсові мережі – це зручний інструмент для класифікації при здійсненні інтелектуального аналізу даних. Однак побудова Байєсових мереж по навчальних даних – це NP- складна проблема. В статті запропоновано евристичний метод побудови Байєсових мереж, оснований на використанні взаємної інформації між всіма вершинами, а як функції оцінки на кожній ітерації алгоритму навчання – використовувати значення описання мінімальної довжини. Для розрахунку похибки навчання запропоновано використовувати формулу структурної різниці. Наведено основні визначення та відповідні ілюстративні приклади.*
Ключові слова: *Байєсова мережа, машинне навчання, принцип описання мінімальної довжини (ОМД), взаємна інформація, евристичний метод навчання.*

Аннотация: *Байесовские сети являются великолепным инструментом для классификации при выполнении интеллектуального анализа данных. Но построение Байесовской сети по обучающим данным является NP-трудной задачей. В статье предлагается эвристический метод построения Байесовских сетей, основанный на использовании обоюдной информации между всеми вершинами, а в качестве оценочной функции, на каждой итерации алгоритма обучения, можно использовать значение описания минимальной длины. Для вычисления ошибки обучения предложено использовать формулу структурной разности. Приведены основные определения и соответствующие иллюстративные примеры.*
Ключевые слова: *Байесовская сеть, машинное обучение, принцип минимальной длины описания (ОМД), обоюдная информация, эвристический метод обучения.*

1. Введение

Многие компании годами накапливают бизнес-информацию, надеясь, что она поможет им в принятии решений. Чем конкретнее информация, тем полезнее она для принятия решений. Интеллектуальный анализ данных (Data Mining) – это технология выявления скрытых взаимосвязей внутри больших баз данных. В основе большинства инструментов интеллектуального анализа данных лежат две технологии: машинное обучение (machine learning) и визуализация (визуальное представление информации). Байесовские сети как раз и объединяют в себе эти две технологии.

Машинное обучение ставит своей задачей выявление закономерностей в эмпирических данных [4, 5]. В противоположность математическому моделированию, изучающему следствия из известных законов, машинное обучение предназначено для воссоздания причин на основе наблюдений – эмпирических данных. Обучающиеся модели должны быть чувствительны к данным благодаря адаптации в процессе обучения своих настроечных параметров с целью наилучшего объяснения всех известных фактов. Однако хорошее качество объяснения имеющихся данных еще не гарантирует соответствующее качество прогнозов. Излишне сложные модели способны адаптироваться не только к типичным закономерностям, но и к случайным событиям, зафиксированным в данной обучающей выборке. Как следствие, такие модели обладают плохой прогнозирующей способностью: большая чувствительность к данным приводит к большому разбросу в прогнозах [4]. Модель в этом случае оказывается неспособной обобщить (усреднить)

данные путем отделения общих закономерностей от случайных флуктуаций. Поэтому ограничение сложности моделей является необходимым элементом теории обучения.

2. Постановка задачи

Задача построения Байесовской сети по заданным обучающим данным является NP-трудной (NP-hard, то есть – это задача нелинейной полиномиальной сложности). Поэтому разработка методов, позволяющих уменьшить вычислительную сложность, является актуальной и востребованной при моделировании процессов различной природы сетями Байеса. Ставится задача разработки эвристического метода построения Байесовских сетей, состоящего из двух этапов. На первом этапе выполняется вычисление значения обоюдной информации между всеми вершинами. На втором выполняется целенаправленный поиск, использующий в качестве оценочной функции оценку минимальной длины (ОМД), основанную на принципе описания, который применяется на каждой итерации алгоритма обучения.

3. Понятие Байесовской сети

Байесовская сеть (БС) – это пара $<G,B>$, в которой первый компонент G является направленным ациклическим графом, соответствующим случайным переменным. Граф записывают как набор условий независимости: каждая переменная независима от ее родителей в G. Вторая компонента пары – B, представляет собой множество параметров, определяющих сеть. Она содержит параметры $\Theta_{x^i \mid pa(X^i)} = P(x^i \mid pa(X^i))$ для каждого возможного значения x^i из X^i и $pa(X^i)$ из $Pa(X^i)$, где $Pa(X^i)$ обозначает набор родителей переменной X^i в G. Каждая переменная X^i в графе G представляется в виде вершины. Если рассмотреть больше чем один граф, то тогда используется обозначение $Pa^G(X^i)$ для определения родителей X^i в графе G. Полная совместная вероятность БС B вычисляется по формуле $P_B(X^1,...,X^N) = \prod_{i=1}^{N} P_B(X^i \mid Pa(X^i))$.

С математической точки зрения БС – это модель представления вероятностных зависимостей, а также отсутствия этих зависимостей. При этом связь $A \to B$ является причинной, когда событие A является причиной возникновения B, то есть, когда есть механизм, в соответствии с которым значение, принятое A, влияет на значение, принятое B. БС называют причинной (каузальной), когда все ее связи являются причинными.

4. Вычислительная сложность задачи построения Байесовской сети

Построение Байесовской сети можно выполнить “в лоб”, простым перебором (exhaustive search) множества всех возможных нециклических моделей, из которых выбрать модель, наиболее адекватно соответствующую обучающим данным. Данная задача является NP-трудной, так как при полном переборе количество всех моделей равняется $3^{\frac{n\cdot(n-1)}{2}} - k_{cycle}$, где n – количество вершин,

k_{cycle} – количество моделей с циклами. Количество всех возможных нециклических моделей можно посчитать при помощи рекуррентной формулы Робинсона, предложенной в 1976 году, [1, 2]:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \cdot C_n^i \cdot 2^{i \cdot (n-i)} \cdot f(n-i),$$

где n – количество вершин, а $f(0) = 1$.

Таблица 1. Таблица зависимости числа моделей без циклов от количества вершин, которые нужно проанализировать при полном переборе моделей

Число вершин	Модели без циклов		Число вершин	Модели без циклов
1	1		6	3,781,503
2	3		7	1,138,779,265
3	25		8	783,702,329,343
4	543		9	1,213,442,454,842,881
5	29,281		10	4,175,098,976,430,598,100

Однако на практике выполнить полный перебор моделей можно только для сетей не более чем с 7 вершинами. При количестве вершин больше 7 выполнить простой перебор не представляется возможным, так как не хватит никаких вычислительных ресурсов. Поэтому предлагается для построения Байесовских сетей использовать эвристический метод. Сначала метод производит вычисление значений обоюдной информации (mutual information) между всеми вершинами, после чего выполняется целенаправленный поиск, использующий в качестве оценочной функции принцип описания минимальной длины (ОМД), который применяется на каждой итерации алгоритма обучения.

5. Значения обоюдной информации (mutual information) между переменными

Для оценки степени зависимости двух произвольных переменных x^i и x^j в работе [3] Шоу и Лью в 1968 году предложили использовать значение обоюдной информации $MI(x^i, x^j)$. Для расчёта предложено следующее выражение:

$$MI(x^i, x^j) = \sum_{x^i, x^j} P(x^i, x^j) \cdot \log\left(\frac{P(x^i, x^j)}{P(x^i) \cdot P(x^j)}\right).$$

По своей сути значение обоюдной информации является аналогом корреляции, но по своему содержанию – это оценка количества информации, содержащейся в переменной x^i о переменной x^j. Значение обоюдной информации принимает неотрицательные значения $MI(x^i, x^j) \geq 0$, а в случае, если вершины x^i и x^j полностью независимы друг от друга, то $MI(x^i, x^j) = 0$, так как $P(x^i, x^j) = P(x^i) \cdot P(x^j)$ и, следовательно,

$$\log\left(\frac{P(x^i, x^j)}{P(x^i) \cdot P(x^j)}\right) = \log\left(\frac{P(x^i) \cdot P(x^j)}{P(x^i) \cdot P(x^j)}\right) = \log(1) = 0.$$

В случае, если Байесовская сеть состоит из N вершин, то для вычисления $MI(x^i, x^j)$ для всех паросочетаний x^i и x^j потребуется выполнить $\frac{N \cdot (N-1)}{2}$ вычисление, при этом $MI(x^i, x^j) = MI(x^j, x^i)$.

6. Принцип описания минимальной длины (ОМД)

Согласно теории кодирования Шеннона, при известном распределении $P(X)$ случайной величины X длина оптимального кода для передачи конкретного значения x по каналу связи стремится к $L(x) = -\log P(x)$. Энтропия источника $S(P) = -\sum_x P(x) \cdot \log P(x)$ является минимальной ожидаемой длиной закодированного сообщения. Любой другой код, основанный на неправильном представлении об источнике сообщений, приведет к большей ожидаемой длине сообщения. Иными словами, чем лучше модель источника, тем компактнее могут быть закодированы данные.

В задаче обучения источником данных является некая неизвестная нам истинная функция распределения $P(D|h_0)$, где $D = \{d_1, ..., d_N\}$ – набор данных, h – гипотеза вероятностного происхождения данных, $L(D|h) = -\log P(D|h)$ – эмпирический риск, аддитивный по числу наблюдений и пропорциональный эмпирической ошибке. Отличие между $P(D|h_0)$ и модельным распределением $P(D|h)$ по мере Кулбака-Леблера определяется как

$$\left|P(D|h) - P(D|h_0)\right| = \sum_D P(D|h_0) \cdot \log \frac{P(D|h_0)}{P(D|h)} = \sum_D P(D|h_0) \cdot \left|L(D|h) - L(D|h_0)\right| \geq 0,$$

то есть оно представляет собой разницу ожидаемой длины кодирования данных с помощью гипотезы и минимально возможной. Эта разница всегда неотрицательна и равна нулю лишь при полном совпадении двух распределений. Иными словами, гипотеза тем лучше, чем короче средняя длина кодирования данных [4]. Принцип ОМД в своей нестрогой и наиболее общей формулировке гласит: среди множества моделей следует выбрать ту, которая позволяет описать данные наиболее коротко, без потери информации [6].

В общем виде задача ОМД выглядит следующим образом. Сначала задается множество обучающих данных $D = \{d_1, ..., d_n\}$, $d_i = \{x_i^{(1)} x_i^{(2)} ... x_i^{(N)}\}$ (нижний индекс – номер наблюдения, а верхний – номер переменной), n-количество наблюдений, каждое наблюдение состоит из N ($N \geq 2$) переменных $X^{(1)}, X^{(2)}, ..., X^{(N)}$, каждая j-я переменная ($j = 1, ..., N$) имеет $A^{(j)} = \{0,1, ..., \alpha^{(j)} - 1\}$ ($\alpha^{(j)} \geq 2$) состояний, каждая структура $g \in G$ БС представляется N множествами предков ($\Pi^{(1)}, ..., \Pi^{(N)}$), то есть для каждой вершины $j = 1, ..., N$, $\Pi^{(j)}$ – это множество родительских вершин, такое что $\Pi^{(j)} \subseteq \{X^{(1)}, ..., X^{(N)}\} \setminus \{X^{(j)}\}$ (вершина не может быть предком самой себе, то есть петли в графе отсутствуют). Тогда ОМД структуры $g \in G$ при

заданной последовательности из n наблюдений $x^n = d_1 d_2 ... d_n$ вычисляется по формуле $L(g, x^n) = H(g, x^n) + \frac{k(g)}{2} \cdot \log(n)$, где $k(g)$ – количество независимых условных вероятностей в сетевой структуре g, а $H(g, x^n)$ – эмпирическая энтропия.

$$H(g, x^n) = \sum_{j \in J} H(j, g, x^n), \quad k(g) = \sum_{j \in J} k(j, g),$$

где ОМД j-й вершины вычисляется по формуле

$$L(j, g, x^n) = H(j, g, x^n) + \frac{k(j, g)}{2} \cdot \log(n);$$

$k(j, g)$ – количество независимых условных вероятностей j-й вершины

$$k(j, g) = (\alpha^{(j)} - 1) \cdot \prod_{k \in \phi(j)} \alpha^k,$$

где $\phi(j) \subseteq \{1,.., j-1, j+1,..., N\}$ – это такое множество, что $\Pi^{(j)} = \{X^{(k)} : k \in \phi^{(j)}\}$.

Эмпирическая энтропия j-й вершины вычисляется по формуле

$$H(j, g, x^n) = \sum_{s \in S(j,g)} \sum_{q \in A^{(j)}} -n[q, s, j, g] \cdot \log \frac{n[q, s, j, g]}{n[s, j, g]},$$

где

$$n(s, j, g) = \sum_{i=1}^{n} I(\pi_i^{(j)} = s); n[q, s, j, g] = \sum_{i=1}^{n} I(x_i = q, \pi_i^{(j)} = s),$$

где $\pi^{(j)} = \Pi^{(j)}$ означает $X^{(k)} = x^{(k)}, \forall k \in \phi^{(j)}$, функция $I(E) = 1$, когда предикат $E = true$, в противном случае $I(E) = 0$.

Простой алгоритм обучения БС с использованием ОМД выглядит следующим образом. По циклу производится перебор всех возможных нециклических сетевых структур. В g^* сохраняется оптимальная сетевая структура. Оптимальной структурой будет та, у которой наименьшее значение функции $L(g, x^n)$.

Простой алгоритм обучения БС с использованием ОМД

1. $g^* \leftarrow g_0 (\in G)$.
2. Для $\forall g \in G - \{g_0\}$, если $L(g, x^n) < L(g^*, x^n)$, то тогда $g^* \leftarrow g$.
3. На выход подаётся g^* в качестве решения.

7. Пример использования метода ОМД

Пусть задан набор обучающих данных из 10 наблюдений для обучения БС, который приведён в табл. 2. В случае полного перебора всех возможных сетевых структур следует рассмотреть 25 структур. После того, как будут рассмотрены все 25 структур, в качестве оптимальной выдаётся структура, изображённая на рис. 1.

Таблица 2. Набор из 10 наблюдений для обучения БС

n	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$		n	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
1	0	1	1		6	0	1	1
2	1	0	0		7	1	0	1
3	0	1	1		8	1	0	0
4	1	0	0		9	0	1	1
5	0	1	1		10	1	1	1

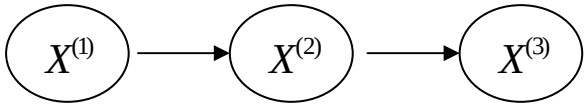


Рис. 1. Оптимальная структура, соответствующая данным из табл. 2

Длина описания этой структуры вычисляется следующим образом. Вершина $X^{(1)}$ не имеет предков, то есть $\Pi^{(1)} = \{\}$. Эмпирическая энтропия вычисляется как $H(j=1,g) = -5 \cdot \log\left(\frac{5}{10}\right) - 5 \cdot \log\left(\frac{5}{10}\right) = 6{,}9315$, а количество независимых условных вероятностей $k(j=1,g) = 2-1 = 1$. Следовательно, длина описания вершины $X^{(1)}$ равняется $L(1,g) = 6{,}9315 + \frac{1}{2} \cdot \log(10) = 8{,}0828$. При вычислении можно использовать логарифм с любой базой. В данном примере используется с базой $e = 2{,}7183$, то есть натуральный логарифм.

Таблица 3. Таблица значений параметров вершины $X^{(1)}$

$X^{(1)}$	$n[q,s,j,g]$	$n[s,j,g]$
0	5	10
1	5	

Вершина $X^{(2)}$ имеет одного предка $X^{(1)}$, то есть $\Pi^{(2)} = \{X^{(1)}\}$. Эмпирическая энтропия вычисляется как

$$H(j=2,g) = \left(-0 \cdot \log\left(\frac{0}{5}\right) - 5 \cdot \log\left(\frac{5}{5}\right)\right) + \left(-4 \cdot \log\left(\frac{4}{5}\right) - 1 \cdot \log\left(\frac{1}{5}\right)\right) = 2{,}502,$$

а количество независимых условных вероятностей $k(j=2,g) = (2-1) \cdot 2 = 2$. Следовательно, длина описания вершины $X^{(2)}$ равняется

$$L(2,g) = 2{,}502 + \frac{2}{2} \cdot \log(10) = 4{,}8046.$$

Таблица 4. Таблица значений параметров вершин $X^{(2)}$ и $X^{(3)}$

$X^{(1)}$	$X^{(2)}$	$n[q,s,j,g]$	$n[s,j,g]$		$X^{(2)}$	$X^{(3)}$	$n[q,s,j,g]$	$n[s,j,g]$
0	0	0	5		0	0	3	4
0	1	5			0	1	1	
1	0	4	5		1	0	0	6
1	1	1			1	1	6	

Вершина $X^{(3)}$ имеет одного предка $X^{(2)}$, то есть $\Pi^{(3)} = \{X^{(2)}\}$. Эмпирическая энтропия вычисляется как

$$H(j=3,g) = \left(-3\cdot\log\left(\frac{3}{4}\right) - 1\cdot\log\left(\frac{1}{4}\right)\right) + \left(-0\cdot\log\left(\frac{0}{6}\right) - 6\cdot\log\left(\frac{6}{6}\right)\right) = 2.2493,$$

а количество независимых условных вероятностей $k(j=3,g) = (2-1)\cdot 2 = 2$. Следовательно, длина описания вершины $X^{(3)}$ равняется

$$L(3,g) = 2{,}2493 + \frac{2}{2}\cdot\log(10) = 4{,}5519.$$

То есть длина описания структуры g, представленной на рис. 1, равна

$$H(g,x^n) = \sum_{j=1}^{3} H(j,g,x^n) = 17{,}4393.$$

О создании и использовании ОМД более подробно можно прочитать в [4, 6, 7, 8].

8. Эвристический метод построения Байесовских сетей

Входные данные. Множество обучающих данных $D = \{d_1,...,d_n\}$, $d_i = \{x_i^{(1)}x_i^{(2)}...x_i^{(N)}\}$ (нижний индекс – номер наблюдения, а верхний – номер переменной), n – количество наблюдений, N – количество вершин (переменных).

Первый этап. Для всех пар вершин вычисляют значения обоюдной информации $Set\_MI = \left\{ MI(x^i, x^j);\ \ \forall i, j \right\}$. После этого элементы множества Set_MI упорядочивают по убыванию $Set\_MI = \{MI(x^{m_1}, x^{m_2}), MI(x^{m_3}, x^{m_4}), MI(x^{m_5}, x^{m_6}),...\}$.

Второй этап. Шаг 1. Из множества значений обоюдной информации Set_MI выбирают первые два максимальные значения $MI(x^{m_1}, x^{m_2})$ и $MI(x^{m_3}, x^{m_4})$. По полученным значениям $MI(x^{m_1}, x^{m_2})$ и $MI(x^{m_3}, x^{m_4})$ строится множество моделей G вида

{($m_1 \to m_2$; $m_3 \to m_4$), ($m_1 \to m_2$; $m_3 \leftarrow m_4$), ($m_1 \leftarrow m_2$; $m_3 \leftarrow m_4$), ($m_1 \leftarrow m_2$; $m_3 \to m_4$), ($m_1 \leftarrow m_2$; m_3 не зависит от m_4), ($m_1 \to m_2$; m_3 не зависит от m_4), (m_1 не зависит от m_2; $m_3 \to m_4$), (m_1 не зависит от m_2; $m_3 \leftarrow m_4$), (m_1 не зависит от m_2; m_3 не зависит от m_4)}. Запись вида $m_i \to m_j$ означает, что вершина x^{m_i} является предком вершины x^{m_j}.

Шаг 2. Затем среди всех моделей множества G осуществляется поиск. В параметре g^* сохраняется оптимальная сетевая структура. Оптимальной структурой будет та, у которой наименьшее значение функции $L(g,x^n)$. $L(g,x^n)$ – ОМД структуры модели при заданной последовательности из n наблюдений $x^n = d_1d_2...d_n$.

1. $g^* \leftarrow g_0 (\in G)$.

2. Для $\forall g \in G - \{g_0\}$, если $L(g, x^n) < L(g^*, x^n)$, то тогда $g^* \leftarrow g$.

3. На выход подаётся g^* в качестве решения.

Шаг 3. После того, как найдена оптимальная структура (структуры) g^* из G, из множества значений обоюдной информации Set_MI выбирают следующее максимальное значение $MI(x^{i\_next_i}, x^{j\_next})$. По полученному значению $MI(x^{i\_next_i}, x^{j\_next})$ и структуре (структурам) g^* строится множество моделей G вида { (g^*; $i\_next \rightarrow j\_next$), ($g^*$; $i\_next \leftarrow j\_next$), (g^*; i_next не зависит от j_next)}. После чего выполняется **шаг 2**.

Условие завершения. Эвристический метод будет выполняться до тех пор, пока не будет проанализировано определённое число элементов множества или все $\frac{N \cdot (N-1)}{2}$ элементы множества Set_MI. Как показывает практика, в большинстве случаев нет смысла выполнять анализ более чем половины (то есть $\frac{N \cdot (N-1)}{4}$) элементов множества Set_MI.

Выходные данные. Оптимальная структура (структуры) g^*.

9. Пример построения сети “Азия” эвристическим методом

В качестве примера используется сеть “Азия” с восемью вершинами. В табл. 5 приведены значения обоюдной информации всех вершин сети (первый этап алгоритма), а в табл. 6 приведён порядок построения БС “Азия” эвристическим методом (второй этап алгоритма). Обучение выполнялось выборкой из 7000 обучающих наблюдений.

Таблица 5. Значения обоюдной информации между всеми вершинами БС “Азия”

№	*MI*	*I*	*j*	№	*MI*	*i*	*j*	№	*MI*	*i*	*j*	№	*MI*	*i*	*j*
1	0,251	7	8	**8**	0,0245	1	8	**15**	0,001227	3	5	**22**	0,00012271	2	5
2	0,136	2	4	**9**	0,0132	4	8	**16**	0,000851	1	6	**23**	0,00006475	5	6
3	0,125	4	6	**10**	0,0101	2	8	**17**	0,000508	2	7	**24**	0,00003950	2	3
4	0,096	2	6	**11**	0,0051	6	8	**18**	0,000381	3	7	**25**	0,00003249	5	7
5	0,048	1	7	**12**	0,0031	1	2	**19**	0,000266	4	5	**26**	0,00001725	5	8
6	0,036	3	4	**13**	0,0028	3	8	**20**	0,000197	1	5	**27**	0,00000303	1	3
7	0,025	3	6	**14**	0,0022	1	4	**21**	0,000128	4	7	**28**	0,00000074	6	7

На рис. 2 приведена структура оригинальной Байесовской сети, по которой генерировались значения.

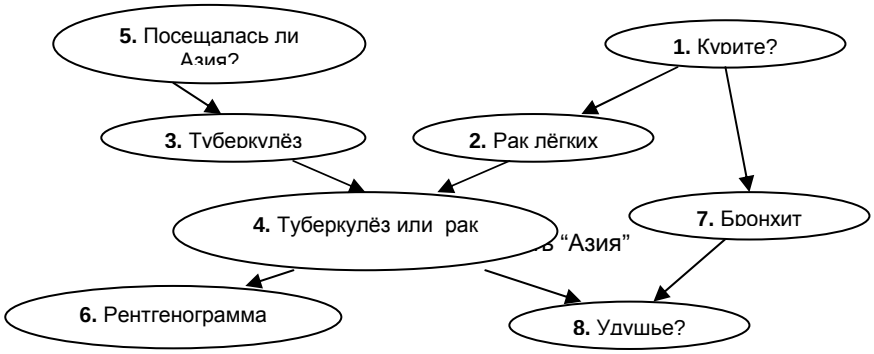
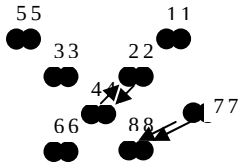
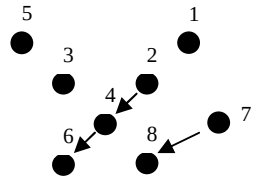
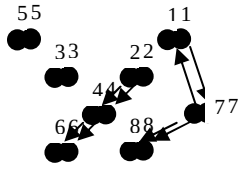
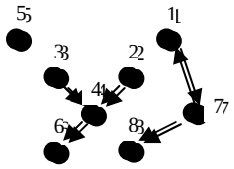
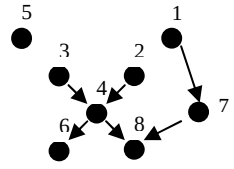
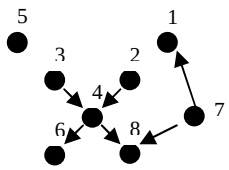
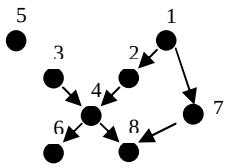
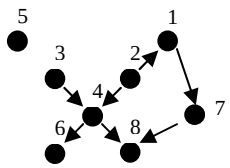
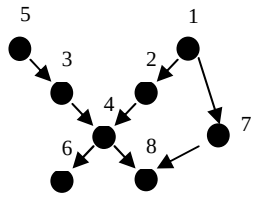


Рис. 2. Оригинальная сеть «Азия»

Таблица 6. Обучение БС “Азия”

Полученная оптимальная структура		Итерация
		На 1-й итерации по первым 2-м строкам *MI(7,8)* и *MI(2,4)* отсортированной матрицы *MI* строится множество моделей из 9 структур
		На 2-й итерации по полученным оптимальным моделям и *MI(4,6)* строится множество моделей из 6 структур
		На 3-й итерации по полученной оптимальной модели и *MI(2,6)* строится множество моделей из 3 структур. В результате получаем ту же оптимальную структуру, что и на предыдущей итерации
		На 4-й итерации по оптимальной модели и *MI(1,7)* строится множество моделей из 3 структур
		На 5-й итерации по оптимальным моделям 4-й итерации и *MI(3,4)* строится множество моделей из 6 структур
		На 6-й итерации по оптимальным моделям 5-й итерации и *MI(3,6)* строится множество моделей из 6 структур. В результате получаем те же оптимальные структуры, что и на предыдущей 5-й итерации
		На 7-й итерации по оптимальным моделям 5-й итерации и *MI(1,8)* строится множество моделей из 6 структур. Результат совпадает с 5-й итерацией
		На 8-й итерации по полученным на предыдущей 7-й итерации моделям и *MI4,8)* строится множество моделей из 6 структур
		На 9-й итерации по оптимальным моделям 8-й итерации и *MI(2,8)* строится множество моделей из 6 структур. В результате получаем те же оптимальные структуры, что и на предыдущей 8-й итерации
		На 10-й итерации по оптимальным моделям 8-й итерации и *MI(6,8)* строится множество моделей из 6 структур. Результат совпадает с 8-й итерацией

Продолжение табл. 6

 	На 11-й итерации по полученным на предыдущей 10-й итерации моделям и *MI(1,2)* строится множество моделей из 6 структур
	На 12-й итерации по *MI(3,8)* строится множество моделей из 6 структур
	На 13-й итерации по *MI(1,4)* строится множество моделей из 6 структур
	На 14-й итерации по полученным на предыдущей 10-й итерации моделям и *MI(3,5)* строится множество моделей из 6 структур
	На 15-й итерации по полученной на 14-й итерации оптимальной структуре и *MI(1,6)* строится множество моделей из 3 структур
	С 15-й по 27-ю итерацию никаких изменений оптимальной структуры, полученной на 14-й, итерации не происходит

Для построения БС “Азия” при простом анализе всех возможных нециклических структур потребуется выполнить оценку 783 702 329 343 моделей. Тогда как эвристический метод на 27-ми итерациях алгоритма выполняет анализ всего лишь 120 структур, причём уже на 14-й итерации, после анализа 81 структуры, метод выдаёт структуру, полностью совпадающую с оригинальной сетью “Азия”. То есть следующие 13 итераций метода не производят никаких изменений, потому что оптимальная структура уже найдена на 14 итерации.

10. Оценка качества обучения БС

Для оценивания качества обучения БС можно использовать учёт количества лишних, отсутствующих и реверсированных дуг в обученной БС по сравнению с оригинальной БС. А в качестве меры ошибки обучения можно использовать структурную разницу (structure difference) или перекрёстную энтропию (cross entropy) между обученной БС и оригинальной БС.

Для вычисления структурной разницы используют формулу **симметрической разницы структур** [9]:

$$\delta = \sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} card\left(\Pi^{(i)}(B) \Delta \Pi^{(i)}(A)\right) = \sum_{i=1}^{n} card\left(\left(\Pi^{(i)}(B) \setminus \Pi^{(i)}(A)\right) \cup \left(\Pi^{(i)}(A) \setminus \Pi^{(i)}(B)\right)\right),$$

где B – обученная БС, A – оригинальная БС, n – количество вершин сети, $\Pi^{(i)}(B)$ – множество предков i – й вершины обученной сети B, $\Pi^{(i)}(A)$ – множество предков i – й вершины оригинальной сети A, $card(\xi)$ – мощность конечного множества ξ, которое определяется как количество элементов, принадлежащих множеству ξ.

Перекрёстная энтропия – это расстояние между распределением обученной БС и оригинальной БС. Пусть $p(v)$ – совместное распределение оригинальной БС, а $q(v)$ – совместное распределение обученной БС. Тогда перекрёстная энтропия вычисляется как [10]

$$H(p,q) = \sum_{v} p(v) \cdot \log \frac{p(v)}{q(v)} = \sum_{j \in J} \sum_{s \in S(j,g)} \sum_{a \in A^{(j)}} p(X^{(j)} = a \left| \Pi^{(j)} = s \right.) \cdot \log \frac{p(X^{(j)} = a \left| \Pi^{(j)} = s \right.)}{q(X^{(j)} = a \left| \Pi^{(j)} = s \right.)}.$$

11. Экспериментальные результаты

Было проведено шесть вычислительных экспериментов. В каждом эксперименте эвристическим методом проводилось обучение сети из 10 вершин выборкой из 2000 обучающих наблюдений. Для оценивания качества обучения используется структурная разница между обученной и оригинальной Байесовской сетью. В табл. 6 показаны результаты шести вычислительных экспериментов. Для каждого эксперимента было выполнено 44 итерации обучения.

Таблица 7. Результаты шести вычислительных экспериментов

Номер вычислительного эксперимента	**№1**	**№2**	**№3**	**№4**	**№5**	**№6**
Общее количество моделей, проанализированных эвристическим методом на всех итерациях	513	178	415	282	550	329
Лишние дуги	1	0	1	2	4	0
Отсутствующие дуги	0	0	0	0	1	0
Реверсированные дуги	3	0	1	1	1	0
Структурная разница между обученной и оригинальной моделями	8	0	3	3	7	0

Как видно из табл. 6, в двух из шести вычислительных экспериментах №2 и №6 обученная сеть полностью совпала с оригинальной БС. В двух из шести экспериментах №3 и №4 ошибка обучения, то есть структурная разница между обученной и оригинальной моделями равняется 3, что для сети из 10 вершин является приемлемой ошибкой. Значительные ошибки обучения получены в экспериментах №1 и №5. Однако для построения сети был выполнен анализ всего лишь 513 и 550 моделей соответственно, на всех 44 итерация, в то время как при простом переборе всех возможных нециклических моделей нужно было бы проанализировать 4 175 098 976 430 598 100 моделей.

12. Выводы

В статье рассмотрена проблема обучения Байесовских сетей. Поскольку обучение БС является NP-трудной задачей, то для уменьшения вычислительной сложности предложен новый эвристический метод построения БС, основанный на использовании оценки обоюдной информации между вершинами и методом ОМД. Данный эвристический метод является итерационным и позволяет значительно уменьшить вычислительную сложность обучения БС.

Алгоритм предложенного эвристического метода подробно рассмотрен на известном примере обучения БС “Азия”, состоящей из 8 вершин. Для обучения понадобилось выполнить анализ 120 структур, тогда как при простом полном переборе нужно проанализировать 783 702 329 343 нециклических структур.

Из результатов проведённых вычислительных экспериментов видно, что в большинстве случаев ошибка обучения эвристическим методом является приемлемой, а экономия вычислительных ресурсов и времени очень значительной. Для оценивания качества обучения сетей использованы формулы структурной разницы и перекрёстной энтропии.

Использование эвристического метода обучения существенно расширяет возможности использования Байесовских сетей при проведении анализа в различных областях человеческой деятельности, особенно там, где приходится работать с большими объёмами информации.

СПИСОК ЛИТЕРАТУРЫ

1. Robinson R.W. Counting unlabeled acyclic digraphs // Proceeding of Fifth Australian on Combinatorial Mathematics. Melbourne. – 1976. – P. 28–43.
2. Leray P., Francois O. BNT structure learn package: documentation and experiments // Technical report, laboratory PSI-INSA Rouen-FRE CNRS 2645. – 2004. – 27 p.
3. Chow C.K., Liu C.N. Approximating discrete probability distributions with dependence trees // IEE Transactions on information theory. –1968. – Vol. IT–14, № 3. – 6 p.
4. Шумский С.А. Байесова регуляризация обучения. Лекции по нейроинформатике. – М.: МИФИ, 2002. – Ч. 2. – 172 с.
5. Бидюк П.И., Терентьев А.Н., Гасанов А.С. Построение и методы обучения Байесовских сетей // Кибернетика и системный анализ. – 2005. – № 4. – С. 133–147.
6. Grunwald P. A Tutorial Introduction to the Minimum Description Length Principle. // Advances in Minimum Description Length: Theory and Applications MIT Press. – Cambridge. – 2005. – 80 p.
7. Suzuki J. Learning Bayesian Belief Networks Based on the MDL Principle: An Efficient Algorithm Using the Branch and Bound Technique // IEICE Trans. on Information and Systems. – 1999. – P. 356–367.
8. Suzuki J. Learning Bayesian Belief Networks based on the Minimum Description length Principle: Basic Properties // IEICE Trans. on Fundamentals. – 1999. – Vol. E82-A № 9. – 9 p.
9. Zheng Y., Kwoh C.K. Improved MDL Score for Learning of Bayesian Networks. Proceedings of the International Conference on Artificial Intelligence in Science and Technology. – 2004. – AISAT. – P. 98–103.
10. Heckerman D., Geiger D., Chickering D. Learning Bayesian Networks: The combination of knowledge and statistical data // Technical report. MSR-TR-94-09. – 1994. – 54 p.