УДК 62-50

А.Н. Терентьев, П.И. Бидюк

МЕТОДЫ ПОСТРОЕНИЯ БАЙЕСОВСКИХ СЕТЕЙ

Введение

Машинное обучение (machine learning) ставит своей задачей выявление закономерностей в эмпирических данных. В противоположность математическому моделированию, изучающему следствия из известных законов, машинное обучение стремится воссоздать причины, наблюдая порожденные ими следствия – эмпирические данные. 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.

Томас Байес, ученик де Муавра, доказал свою знаменитую теорему где-то около 1750 года при рассмотрении задачи, "обратной проблеме Бернулли". Опубликована работа Байеса была лишь после его смерти в 1764 году. Современный вид, как и своё имя, теорема приобрела в трудах Лапласа в 1819 году.

Несмотря на свою простоту и очевидность, она стала настоящим яблоком раздора в математической статистике. Противники байесовской статистики считают её бесполезной в силу произвольности выбора априорных вероятностей. На практике байесовское сравнение моделей в 1939 году начал применять кембриджский геофизик сэр Джеффрис.

В работе [6] Купер и Гершкович (Cooper and Herskovits) предлагают базовый метод КГ для обучения Байесовских сетей (БС). В этой же работе они выполняют ряд модификаций улучшающих метод, в результате чего получается общеизвестный и часто используемый метод К2.

В работах [1 и 2] Джо Сузуки (Joe Suzuki) исследует проблему обучения БС при помощи метода описания минимальной длины (ОМД). В работе [1] Сузуки вместе с ОМД предлагает использовать метод ветвей и границ, полученный метод выполняет более качественное обучение, по сравнению с простым методом ОМД. В работе [2] Сузуки описывает отличие ОМД и КГ методов, а так же предлагается модификация метода Шоу

 ISSN 1562-9945

и Лью (Show and Liu). Модифицированный метод, по сравнению с ОМД методами, показывает более лучшие экспериментальные результаты.

Звенг Янь и Квог Чи (Zheng Yun and Kwoh Chee) в работе [7] предлагают модификацию ОМД метода путём уменьшения длины описания таблицы условных вероятностей и описания всей длины модели. Экспериментальные результаты показывают, что модифицированный метод выполняет обучение лучше.

Теоретический обзор ОМД метода, начиная с его зарождения и заканчивая различными улучшениями и модификациями, выполнен голландским исследователем Питером Грюнвальдом в работе [3].

Несмотря на то, что байесовским сетям уделяется много внимания в зарубежной литературе, принципы их построения и использования еще недостаточно освещены в отечественных публикациях, что существенно затрудняет их понимание и применение.

Постановка задачи.

Для множества связанных событий $X^{(i)},\ i = 1, \ldots, N$ задается множество обучающих данных $D = \{d_1, \ldots, d_n\}$, $d_i = \{x_i^{(1)} x_i^{(2)} \ldots x_i^{(N)}\}$ (нижний индекс - номер наблюдения, а верхний - номер переменной), n-количество наблюдений, каждое наблюдение состоит из N ($N \geq 2$) переменных, каждая j-я переменная ($j = 1, \ldots, N$) имеет $A^{(j)} = \{0, 1, \ldots, \alpha^{(j)} - 1\}$ ($\alpha^{(j)} \geq 2$) состояний. На основе заданной обучающей выборки нужно построить связывающий множества событий $X_i,\ i = 1, \ldots, N$ ациклический граф. При этом каждая структура $g \in G$ БС представляется N множествами предков ($\Pi^{(1)}, \ldots, \Pi^{(N)}$), то есть для каждой вершины $j = 1, \ldots, N$, $\Pi^{(j)}$ – это множество родительских вершин, такое что $\Pi^{(j)} \subseteq \{X^{(1)}, \ldots, X^{(N)}\} \backslash \{X^{(j)}\}$. Необходимо выполнить исследование некоторых методов построения таких байесовских сетей, а так же привести иллюстрации возможностей их использования на практике. При этом предполагается, что на события $X^{(i)},\ i = 1, \ldots, N$ влияют неопределенности различного характера и природы, а также имеются данные, описывающие эти события.

Понятие байесовской сети

Байесовская сеть (БС) – это пара $< G, B >$, в которой первый компонент G, является направленным ациклическим графом, соответствующий случайным переменным. С байесовскими сетями связано более сложное понятие независимости, которое учитывает направленность дуг. 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 $\Theta_{x^i | \, pa(X^i)} = P(x^i \, | pa(X^i))$ 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

$$P_B(X^1,\ldots,X^N)=\prod_{i=1}^{N} P_B(X^i\,\big|\,Pa(X^i)).$$

С математической точки зрения БС – это модель для 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.

Процесс построения нециклического графа, соответствующего переменным, называется обучением БС, потому что по заданной (обучающей) выборке выполняется вычисление наиболее подходящей сети. Данная задача является NP-трудной (NP-hard), так как при полном переборе (exhaustive search) количество всех моделей равняется $3^{\frac{n\cdot(n-1)}{2}} - k_{cycle}$, где n – количество вершин, k_{cycle} – количество моделей с циклами.

Таблица 1.

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

Всего вершин	Всего моделей	Модели с циклами	Модели без циклов
2	3	0	3
3	27	2	25
4	729	186	543
5	59049	29768	29281

Существует широкий набор различных методов для анализа моделей, но в данной статье будут подробно рассмотрены два метода: принцип минимальной длины описания (minimum description length – MDL) [1, 2] и метод предложенный Купером и Гершковичем (Cooper and Herskovits procedure) [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 $S(P) = -\sum\limits_{x} P(x)\cdot\log P(x)$ 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 наша модель источника, тем компактнее могут быть закодированы данные.

В задаче обучения источником данных является некая неизвестная нам истинная функция распределения $P(D\,|h_0)$, $D = \{d_1, \ldots, d_N\}$ – 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

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

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) формулируется в рамках Колмогоровской алгоритмической теории информации. По сути ОМД является обращением задачи оптимального кодирования. Если в теории Шеннона из знания модели источника сообщений извлекают знание оптимальных кодов, то в ОМД делается наоборот: из поиска оптимальных кодов, оптимального представления данных, извлекается модель источника. Такие методы как анализ главных компонент (АГК), факторный анализ (ФА), анализ независимых компонент (АНК) могут рассматриваться в качестве частных примеров применения с определенными упрощениями принципа ОМД. Данный принцип в своей нестрогой и наиболее общей формулировке гласит: среди множества моделей следует выбрать ту, которая позволяет описать данные наиболее коротко без потери информации [3].

В общем виде задача ОМД выглядит следующим образом. Сначала задается множество обучающих данных $D = \{d_1, \ldots, d_n\}$, $d_i = \{x_i^{(1)} x_i^{(2)} \ldots x_i^{(N)}\}$ (нижний индекс – номер наблюдения, а верхний – номер переменной), n – количество наблюдений, каждое наблюдение состоит из N ($N \geq 2$) переменных $X^{(1)}, X^{(2)}, \ldots, X^{(N)}$, каждая j-я переменная ($j = 1, \ldots, N$) имеет $A^{(j)} = \{0, 1, \ldots, \alpha^{(j)} - 1\}$ ($\alpha^{(j)} \geq 2$) состояний, каждая структура $g \in G$ БС представляется N множествами предков ($\Pi^{(1)}, \ldots, \Pi^{(N)}$), то есть для каждой вершины $j = 1, \ldots, N$, $\Pi^{(j)}$ – это множество родительских вершин, такое что $\Pi^{(j)} \subseteq \{X^{(1)}, \ldots, X^{(N)}\} \backslash \{X^{(j)}\}$ (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 $x^n = d_1 d_2 \ldots d_n$ 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.

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

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 $\phi(j) \subseteq \{1,\ldots,j-1,j+1,\ldots,N\}$ 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 $I(E) = 0$.

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

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.

Пример использование метода ОМД

Пусть у нас задан набор обучающих данных из 10 наблюдений для обучения БС, который приведён в таблице 2.

В случае полного перебора всех возможных сетевых структур нужно будет рассмотреть 25 структур ($3^{\frac{N(N-1)}{2}} = 3^{\frac{3(3-1)}{2}} = 27$ структур, но 2 структуры не рассматриваются, потому что в них присутствуют циклы). После того как будут рассмотрены все 25 структур, в качестве оптимальной будет выдана структура изображённая на рисунке 1.

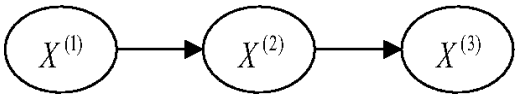

Рис. 1 – Оптимальная структура, соответствующая данным из таблицы 2.

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

Таблица 2.

Набор из 10 наблюдений для обучения БС.

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

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 для вычисления длины описания 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

$$H(j=2,g)=\left(-0\cdot\log\left(\tfrac{0}{5}\right)-5\cdot\log\left(\tfrac{5}{5}\right)\right)+\\+\left(-4\cdot\log\left(\tfrac{4}{5}\right)-1\cdot\log\left(\tfrac{1}{5}\right)\right)=2.502\quad,$$

а количество независимых условных вероятностей $k(j=2,g)=(2-1)\cdot 2=2$. Следовательно длина описания вершины $X^{(2)}$ равняется $L(2,g)=2.502+\frac{2}{2}\cdot\log(10)=4.8046$.

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

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

а количество независимых условных вероятностей $k(j=3,g)=(2-1)\cdot 2=2$. Следовательно длина описания вершины $X^{(3)}$ равняется $L(3,g)=2.2493+\frac{2}{2}\cdot\log(10)=4.5519$.

Таблица 4.

Таблица значений параметров для вычисления длины описания вершины $X^{(2)}$

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

Таблица 5.

Таблица значений параметров для вычисления длины описания вершины $X^{(3)}$

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

То есть длина описания структуры g представленной на рисунке 1 равна $H(g,x^n)=\sum\limits_{j=1}^{3}H(j,g,x^n)=17.4393$.

О создании и использовании ОМД более подробно можно прочитать в [1, 2, 3, 4].

Метод Купера и Гершковича (КГ)

Если в ОМД ищут структуру с минимальной длинной описания, то метод КГ заключается в том, что нужно найти структуру с максимальным значение функции $P(g,x^n)$. В литературе модернизированный метод КГ известен так же как К2 метод.

$$P(g,x^n)=P(g)\cdot\prod_{j\in J}\left(\prod_{s\in S(j,g)}\frac{\left(\alpha^{(j)}-1\right)!\cdot\prod\limits_{q\in A^{(j)}}(n[a,s,j,g]\,!)}{(n[s,j,g]+\alpha^{(j)}-1)\,!}\right)$$

Алгоритм обучения БС с использованием метода КГ выглядит следующим образом, по циклу производится перебор всех возможных не циклических сетевых структур. В g^* сохраняется оптимальная сетевая структура. Оптимальной структурой будет та, у которой будет наибольшее значение функции $P(g,x^n)$.

1. $g^*\leftarrow g_0(\in G)$;
2. для $\forall g\in G-\{g_0\}$ если $P(g,x^n)>P(g^*,x^n)$ то тогда $g^*\leftarrow g$;
3. на выход подаётся g^* в качестве решения.

В качестве примера посчитаем значение функции КГ для структуры показанной на рисунке 1, на основе 10 обучающих данных приведённых в таблице 2.

$$P(1,g,x^n)=\frac{(2-1)!\cdot 5!\cdot 5!}{(10+2-1)!}=0.00036;$$

$$P(2,g,x^n)=\frac{(2-1)!\cdot 0!\cdot 5!}{(5+2-1)!}\cdot\frac{(2-1)!\cdot 4!\cdot 1!}{(5+2-1)!}=0.0056;$$

$$P(3,g,x^n)=\frac{(2-1)!\cdot 3!\cdot 1!}{(4+2-1)!}\cdot\frac{(2-1)!\cdot 0!\cdot 6!}{(6+2-1)!}=0.0071;$$

$$P(g,x^n)=\frac{1}{25}\cdot\prod_{j\in J}P(j,g,x^n)=5.7254\cdot 10^{-10}$$

Более подробная информация о применении и выводе метода КГ дана в [6].

Уменьшение вычислительной сложности

При полном переборе всех возможных структур требуется выполнить анализ $3^{\frac{n\cdot(n-1)}{2}}-k_{cycle}$ числа моделей, где n – количество вершин, k_{cycle} – количество моделей с циклами.

Для того, что бы уменьшить количество структур, которые нужно проанализировать, можно сделать допущение о том, что вершины упорядочены $X^{(1)}<X^{(2)}<\ldots<X^{(n)}$. Запись $X^{(i)}<X^{(j)}$ означает, что вершина $X^{(i)}$ предшествует вершине $X^{(j)}$, то есть вершина $X^{(j)}$ не может быть предком вершины $X^{(i)}$, дуги могут идти только из $X^{(i)}$ в $X^{(j)}$. Используя предварительную упорядоченность вершин, количество структур необходимых для рассмотрения уменьшается до $2^{\frac{n(n-1)}{2}}$, где n – количество вершин. Это очень существенное допущение, которое позволяет экономить вычислительные мощности, при $n=10$ число структур уменьшается в 10^7 раз. Но при этом требуется вмешательство эксперта, который разбирается в анализируемой предметной области и будет выполнять предварительную упорядоченность вершин. Либо можно упорядочивать вершины на основе анализа характеристик обучающих данных, например можно использовать корреляцию атрибутов вершин [2] или табу поиск (taboo search) [9].

Но даже при введении предварительной упорядоченности вершин объём вычислений остаётся огромным, например при $n=8$ вершин, потребуется выполнить анализ $2^{\frac{8*(8-1)}{2}}=268.435.456$ структур. При использовании метода КГ возникает проблема вычисления факториала, приведём тривиальный пример, когда у нас две вершины в структуре $X^{(1)}$ и $X^{(2)}$, а множество обучающих примеров состоит из миллиона записей $D=\{d^{(1)},\ldots,d^{(1.000.000)}\}$, при вычислении $P(g,x^n)$ потребуется посчитать факториал вида $(n[s,j,g]+\alpha^{(j)}-1)!=(1.000.000+\alpha^{(j)}-1)!$, в то время

как такие пакеты как Mat Lab и MathCAD вычисляют факториалы не более 170!.

Способы оценивания качества построения 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 enthropy), между обученной БС и оригинальной БС.

Для вычисления структурной разницы используют формулу симметрической разницы структур [7]:

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

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 [8]:

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

Практические результаты

В данной статье результаты работы ОМД метода сравниваются с реализованным в программе BayesiaLab (http://www.bayesia.com) методом для обучения БС. Реализованный в программе BayesiaLab метод называется табу упорядоченным обучением (taboo order learning). Табу упорядоченным обучением (ТУО) заключается в том, что сначала выполняется упорядочивание вершин БС методом табу поиска (taboo search), после чего выполняется поиск наилучшей сети.

В качестве примера используется сеть “Азия” с восемью вершинами. В таблице 6 приведены численные результаты эксперимента обучения БС “Азия” ТУО и ОМД методами. В первой серии экспериментов выполнялось обучение выборкой из 100 обучающих наблюдений, во второй серии выборкой из 1000 наблюдений, в третьей серии выборкой из 7000 наблюдений. Как видно из таблицы 6 по мере увеличения числа обучающих

данных ошибка δ между обученной и оригинальной БС уменьшается, в качестве меры ошибки использовалась структурная разница между обученной БС и оригинальной БС. На рисунке 2 нарисована структура оригинальной БС (по которой генерировались значения), на рисунках 3 и 4 представлены структуры обученных сетей ТУО и ОМД методами соответственно, на основе обучающей выборки из 7000 наблюдений.

Таблица 6.

Таблица зависимости ошибки δ от размера обучающей выборки, при обучении БС ТУО и ОМД методами

метод	Nобучающих записей	Лишние дуги	Отсутствующие дуги	Реверсированные дуги	Структурная разность δ
ТУО	$N = 100$	0	3	2	7
ОМД	$N = 100$	0	2	2	6
ТУО	$N = 1000$	0	1	1	3
ОМД	$N = 1000$	0	1	2	5
ТУО	$N = 7000$	0	0	1	2
ОМД	$N = 7000$	0	0	1	2

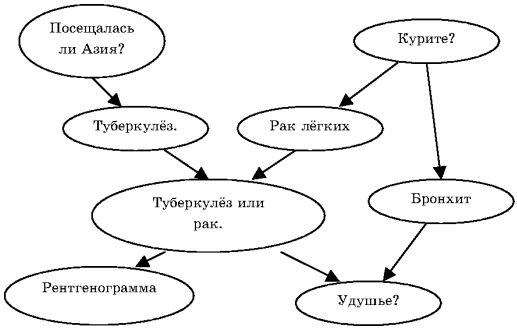

Рис. 2 – Оригинальная сеть “Азия”

Выводы

В статье рассмотрена проблема обучения Байесовских сетей. Основными методами обучения на сегодня являются метод описания минимальной длины, метод Купера и Гершковича, а также различные модернизации этих методов. Подробно на примерах рассмотрены метод минимальной длины описания, а также метод Купера и Гершковича. Поскольку обучение БС является NP-трудной задачей, то для уменьшения вычислительной сложности предлагается использовать предварительную упорядоченность вершин. Для оценивания качества обучения сетей

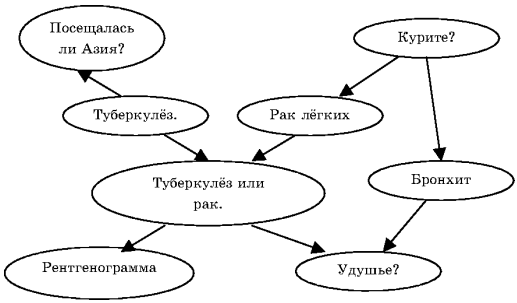

Рис. 3 – Обученная сеть “Азия” ТУО методом, набором обучающих данных из 7000 наблюдений

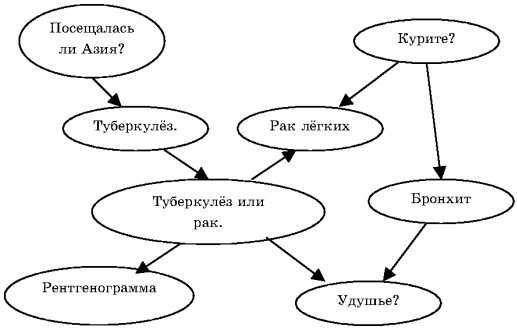

Рис. 4 – Обученная сеть “Азия” ОМД методом, набором обучающих данных из 7000 наблюдений

 ISSN 1562-9945

рассмотрены формулы структурной разницы и перекрёстной энтропии. Приведены практические результаты вычислительных экспериментов, в которых сравниваются ОМД метод и ТУО. Так, для примера сети "Азия" структурная разность принимает минимальное значение при максимальном числе обучающих данных ($N = 7000$). Дальнейшее увеличение объема выборки в данном примере оказалось невозможным из-за ограниченности базы данных.

При моделировании необходимо учитывать вычислительные ограничения моделирующих систем, связанные с конечной длиной разрядной сетки. Например, в 32 разрядных программах Mat Lab и MathCad , возникает проблема вычисления факториалов при $n > 170$, что приводит к затруднениям использования метода КГ. В дальнейшем предполагается применение БС для решения конкретных задач с использованием оригинальных методов обучения, которые имеют приемлемые вычислительные характеристики.

Литература

1. 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. pages Feb. 1999, 356-367 p.
2. Suzuki J. Learning Bayesian Belief Networks based on the Minimum Description length Principle: Basic Properties. // IEICE Trans. on Fundamentals, Vol. E82-A NO 9, September 1999, 9 p.
3. Grunwald P. A Tutorial Introduction to the Minimum Description Length Principle. // Advances in Minimum Description Length: Theory and Applications MIT Press , Cambridge, MA, USA, 2005, 80 p.
4. Шумский С.А. Байесова регуляризация обучения. Лекции по нейроинформатике. Часть 2. – М.: МИФИ, 2002. – 172 с.
5. Vapnik V. The nature of statistical learning theory. Springer, 1995, 188 p.
6. Cooper G. F., Herskovits E. A bayesian method for the induction of probabilistic networks from data. Knowledge Systems Laboratory// Report KSL-91-02, November 1993, 43 p.
7. Zheng Y. and Kwoh C.K. Improved MDL Score for Learning of Bayesian Networks. Proceedings of the International Conference on Artificial Intelligence in Science and Technology, AISAT 2004, 98-103 p.
8. Heckerman D., Geiger D., Chickering D. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Technical Report MSR-TR-94-09, march 1994, 54 p.
9. Park J. and Darwiche A. Complexity Results and Approximation Strategies for MAP Explanations. Journal of Artificial Intelligence Research 21, 2004, 101-133 p.