

БАЙЕСОВСКАЯ СЕТЬ – ИНСТРУМЕНТ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Введение

В мире накапливаются большие объемы информации, которые требуют надлежащей обработки и принятия решений на основе результатов этой обработки. Методы интеллектуального анализа данных (ИАД) предоставляют возможность автоматического поиска закономерностей, характерных для многомерных 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.

В настоящее время технологии ИАД быстро развиваются, а области их применения расширяются. Это связано с совершенствованием методов анализа процессов различной природы и вычислительных процедур, реализующих ИАД. ИАД активно используют такие крупные корпорации как American Express, Lockheed, Google и многие другие. Особенно интенсивно методы ИАД применяются в бизнес-приложениях аналитиками и руководителями компаний. Это обусловлено тем, что успешное развитие любого предприятия напрямую зависит от его способности адекватно и оперативно реагировать на изменение внешней среды, а также от умения прогнозировать результаты тех или иных воздействий. Так, в отчете Ассоциации американских банкиров отмечается, что из 100 крупнейших банков США 45 банков уже внедрили системы интеллектуального анализа данных и еще около 50 банков начали реализацию подобных проектов или планируют это сделать в ближайшее время.

При использовании байесовских сетей (БС) в качестве инструмента ИАД необходимо решать две математические задачи: (1) построения структуры БС и (2) формирование вероятностного вывода. Задача построения БС по заданным обучающим данным является NP-трудной, то есть задачей нелинейной полиномиальной сложности. Количество всех возможных нециклических моделей, которые нужно проанализировать вычисляется по рекуррентной формуле Робинсона, предложенной в 1976 году [1, 2]:

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

где n – количество вершин, а $f(0)=1$. Однако на практике выполнить полный перебор моделей можно только для сетей не более чем с 7 вершинами (узлами), потому что в противном случае не хватает никаких вычислительных ресурсов [8, 9].

Задача формирования вероятностного вывода в БС является важной и сложной и относится к классу задач принятия решений. Один из методов такого формирования предложен в работе [3]. Однако для реализации этого метода необходимо привести структуру БС к виду объединённого дерева (junction tree), и только после этого можно использовать алгоритм вероятностного вывода в объединённом дереве, который основывается на прохождении сообщений λ и π по дереву.

В работе [4] предложен метод поглощающего исключения (bucket elimination). Для его использования необходимо обязательное наличие упорядоченного множества вершин, что является сложной вычислительной задачей. Более подробно анализ проблемы и обзор методов формирования вероятностного вывода в БС рассмотрен в работах [5 – 7].

Очевидно, что существующие методы построения БС и формирования вывода, требуют трудоёмких вычислений. Поэтому разработка методов, позволяющих уменьшить вычислительную сложность, является актуальной и востребованной при моделировании процессов различной природы сетями Байеса.

Постановка задачи

1. Ставится задача разработки эвристического метода построения БС, состоящего из двух этапов [10]. 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.

2. Также ставится задача разработки метода вероятностного вывода в БС с использованием двух шагов. На первом шаге выполняется вычисление матрицы эмпирических значений совместного распределения вероятностей всей сети. На втором шаге проводится вычисление значений вероятностей всех возможных состояний неинстанцированных вершин.

Понятие байесовской сети

Байесовская сеть представляет пару $\langle G, B \rangle$, в которой первая компонента G – это направленный нециклический граф, который соответствует случайным переменным и

записывается как набор условий независимости: каждая переменная независима от ее родителей в G . Вторая компонента пары B – это множество параметров, определяющих сеть. Компонента содержит параметры $\Theta_{X^{(i)}|pa(X^{(i)})} = P(X^{(i)}|pa(X^{(i)}))$ для каждого возможного значения $x^{(i)} \in X^{(i)}$ и $pa(X^{(i)}) \in Pa(X^{(i)})$, где $Pa(X^{(i)})$ обозначает набор родителей переменной $X^{(i)} \in G$. Каждая переменная $X^{(i)} \in G$ представляется в виде вершины. Если рассматривают более одного графа, то для определения родителей переменной $X^{(i)}$ в графе G используют обозначение $Pa^G(X^{(i)})$. Полная совместная вероятность БС вычисляется по формуле $P_B(X^{(1)}, ..., X^{(N)}) = \prod_{i=1}^N P_B(X^{(i)}|Pa(X^{(i)}))$.

Эвристический метод построения Байесовских сетей

Входные данные: множество обучающих экспериментальных 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).

1 этап. Для всех пар вершин вычисляют значения обоюдной информации, $Set_MI = \{MI(x^i, x^j); \forall i, j\}$, по формуле:

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

После этого элементы множества Set_MI упорядочивают по убыванию $Set_MI = \{MI(x^{m_1}, x^{m_2}), MI(x^{m_3}, x^{m_4}), MI(x^{m_5}, x^{m_6}), ... \}$.

2 этап.

Шаг 1. Из множества значений обоюдной информации Set_MI выбирают первые два максимальных 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:

$\{ (m_1 \rightarrow m_2; m_3 \rightarrow m_4), (m_1 \rightarrow m_2; m_3 \leftarrow m_4), (m_1 \leftarrow m_2; m_3 \leftarrow m_4), (m_1 \leftarrow m_2; m_3 \rightarrow m_4), (m_1 \leftarrow m_2; m_3 \text{ не зависит от } m_4), (m_1 \rightarrow m_2; m_3 \text{ не зависит от } m_4), (m_1 \text{ не зависит от } m_2; m_3 \rightarrow m_4), (m_1 \text{ не зависит от } m_2; m_3 \leftarrow m_4), (m_1 \text{ не зависит от } m_2; m_3 \text{ не зависит от } m_4) \}$.

Запись вида $m_i \rightarrow m_j$ означает, что вершина x^{m_i} является предком вершины x^{m_j} .

Шаг 2. Затем осуществляется поиск, среди всех моделей множества G . В параметре g^* сохраняется оптимальная сетевая структура. Оптимальной структурой будет та, у которой

будет наименьшее значение функции $L(g, x^n)$. $L(g, x^n)$ – ОМД структуры $g \in G$ при заданной последовательности из n наблюдений $x^n = d_1 d_2 \dots d_n$ 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 $\phi(j) \subseteq \{1, \dots, j-1, j+1, \dots, N\}$ – 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

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

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

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 $g^* \leftarrow g$;
3. на выход подаётся g^* в качестве решения.

Шаг 3. После того как найдены оптимальные структура (структуры) g^* из G . Из множества значений обоудной информации Set_MI выбирают следующее максимальное значение $MI(x^{i_next}, x^{j_next})$. По полученному значению $MI(x^{i_next}, x^{j_next})$ и структуре (структурам) g^* строится множество моделей G вида: $\{(g^*; i_next \rightarrow j_next), (g^*; i_next \leftarrow j_next), (g^*; i_next \text{ не зависит от } j_next)\}$. 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 $\frac{N \cdot (N-1)}{4}$) элементы множества Set_MI .

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

Алгоритм вероятностного вывода в БС на основе обучающих данных

Входные данные, необходимые для алгоритма формирования вывода.

1. Множество обучающих данных $D = \{d_1, \dots, d_n\}$, где $d_i = \{x_i^{(1)} x_i^{(2)} \dots x_i^{(N)}\}$ (нижний индекс – номер наблюдения, а верхний - номер переменной), n – количество наблюдений. Каждое наблюдение состоит из N ($N \geq 2$) переменных $X^{(1)}, X^{(2)}, \dots, X^{(N)}$, каждая j -я переменная ($j = 1, \dots, N$) имеет $A^{(j)} = \{0, 1, \dots, \alpha^{(j)} - 1\}$ ($\alpha^{(j)} \geq 2$) состояний.

2. Структура БС g представлена множеством из N предков ($\Pi^{(1)}, \dots, \Pi^{(N)}$), т.е. для каждой вершины $j = 1, \dots, N$; $\Pi^{(j)}$ – множество родительских вершин, при этом $\Pi^{(j)} \subseteq \{X^{(1)}, \dots, X^{(N)}\} \setminus \{X^{(j)}\}$ (вершина не может быть предком для самой себя, то есть петли в графе должны отсутствовать).

3. Множество инстанцированных вершин $\{X^{(P_1)} = x^{(P_1)}, \dots, X^{(P_v)} = x^{(P_v)}\}$, т.е. вершин, находящихся в некотором определенном состоянии с единичной вероятностью.

Если множество инстанцированных вершин пустое, то нужно использовать классический вероятностный вывод.

Алгоритм формирования вывода.

Шаг 1. По множеству обучающих данных вычисляется матрица эмпирических значений совместного распределения вероятностей всей сети $P(X^{(1)}, \dots, X^{(N)})$. По формуле

$$P_{matrix}(X^{(1)} = x^{(1)}, \dots, X^{(N)} = x^{(N)}) = \frac{n[X^{(1)} = x^{(1)}, \dots, X^{(N)} = x^{(N)}]}{n}, \text{ где } n - \text{ количество обучающих}$$

наблюдений, $x^{(j)} \in A^{(j)}$, а $n[X^{(1)} = x^{(1)}, \dots, X^{(N)} = x^{(N)}] = \sum_{j=1}^n I(X^{(1)} = x_j^{(1)}, \dots, X^{(N)} = x_j^{(N)})$, где

функция $I(E) = 1$ когда предикат $E = true$, в противном случае $I(E) = 0$.

Шаг 2. Перебираем последовательно все вершины БС. Если вершина не является инстанцированной, то нужно вычислить значения вероятностей всех возможных состояний этой вершины. Для этого делается последовательный перебор всех строк матрицы эмпирических значений совместного распределения вероятностей всей сети. Если значения вершин строки совпадают со значениями инстанцированных вершин и состоянием

анализируемой вершины, то соответствующее значение $P_{matrix}(X^{(1)}, ..., X^{(N)})$ прибавляется к значению вероятности соответствующего состояния анализируемой вершины. После этого нормируются значения вероятностей состояний анализируемой вершины.

Представим алгоритм вычисления значений вероятностей всех возможных состояний неинстанцированных вершин:

```

for j=1 to N if  $X^j \notin \{X^{(P_1)}, ..., X^{(P_v)}\}$  then
  begin
    sum=0;
     $\forall x^{(j)} \in A^{(j)}$  do
      begin
        for k=1 to last_string_matrix do
          begin
            if  $(X_{matrix}^{(P_1)} = x^{(P_1)})$  and ... and  $(X_{matrix}^{(P_v)} = x^{(P_v)})$  and  $(X_{matrix}^{(j)} = x^{(j)})$  then
              begin
                 $P(X^{(j)} = x^{(j)}) = P(X^{(j)} = x^{(j)}) + P_{matrix}(X_{matrix}^{(1)}, ..., X_{matrix}^{(N)});$ 
              end;
            end;
          sum = sum +  $P(X^{(j)} = x^{(j)})$ ;
        end;
       $\forall x^{(j)} \in A^{(j)}$  do
        begin
           $P(X^{(j)} = x^{(j)}) = \frac{P(X^{(j)} = x^{(j)})}{sum};$ 
        end;
      end;
  end;

```

Выходные данные.

Выходными данными являются значения вероятностей всех возможных состояний всех неинстанцированных вершин.

Примеры практического использования БС

Маркетинговый анализ. Для разработки эффективного маркетингового плана, можно использовать не только, такие факторы как стоимость товара, затраты на продвижение продукции и рекламу. Можно и нужно учитывать персональную информацию о потенциальном клиенте, такую как его привычки, возраст, доход и прочее. Решение подобной задачи позволяет настраивать маркетинг и стратегию продвижения товаров для каждой потребительской группы отдельно. Это позволяет увеличивать продажи и обеспечивает лояльное отношение к торговой марке.

Пример 1. Фирма реализует несколько видов автомобильной продукции. У каждого продукта есть своя потребительская группа. В базе данных компании есть информация о 6400 клиентах фирмы, которые уже приобрели товары. По этим данным построена БС, на рис. 1, которая показывает зависимости между сведениями о клиенте и товаром, который он приобретёт.

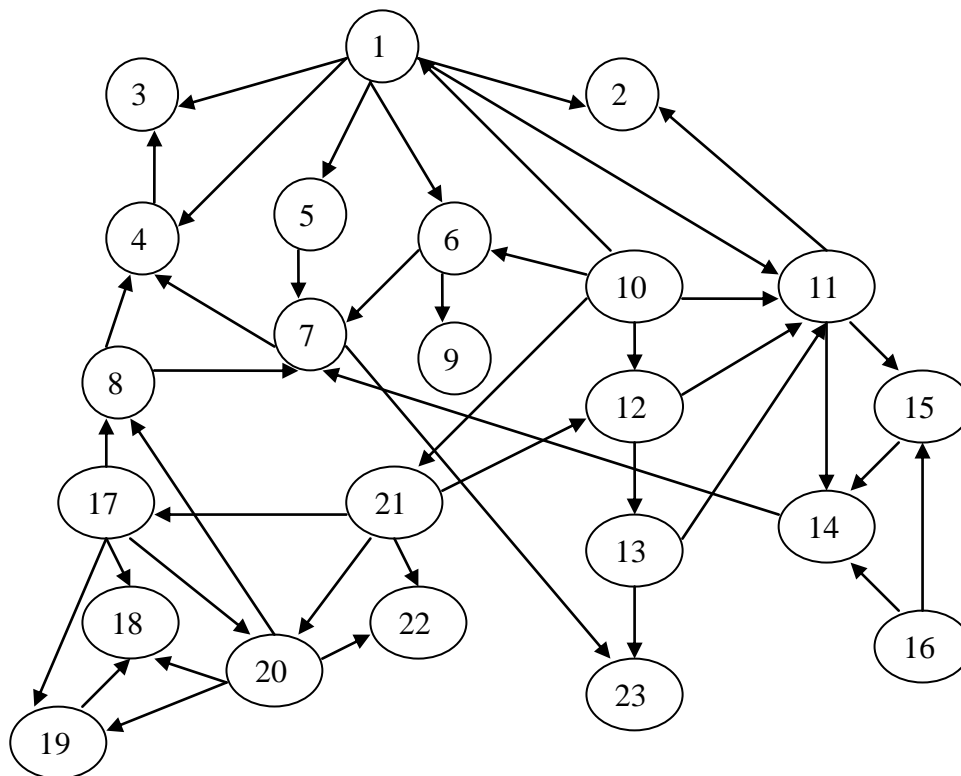


Рис 1

1 – возраст клиента; 2 – пенсионер; 3 – количество детей; 4 – семейное положение; 5 – сколько лет проживает по адресу, указанному в анкете; 6 – стаж работы на текущем рабочем месте; 7 – пользуется информационными новостями; 8 – телефон; 9 – пейджер; 10 – общий трудовой стаж; 11 – доход; 12 – цена автомобиля; 13 – категория автомобиля; 14 – телевизор; 15 – CD-проигрыватель; 16 – видеомэгнитофон; 17 – беспроводная связь; 18 – факс; 19 – органайзер; 20 – компьютер; 21 – уровень образования; 22 – пользуется Интернетом; 23 – потребительская категория.

Анализ результатов маркетинговых исследований. Чтобы оценить реакцию покупателей на политику компании в области распространения продукции, ценообразования, а также на характеристики самой продукции, необходимо, наряду с анализом продаж, проводить опросы покупателей. Это позволяет совершенствовать процесс принятия решений не только по ценам и характеристикам выпускаемой продукции, но учитывая потребности клиентов, работать на перспективу, предлагая своевременно новые услуги.

Пример 2. Телекоммуникационная компания провела опрос 1000 своих клиентов относительно потребляемых ими услуг. На рис. 2 представлена БС, иллюстрирующая связь между личными данными о клиенте и пакетом услуг, предоставляемых клиенту компанией. Подобная модель позволяет оценить реальные потребности клиентов и перспективы их развития.

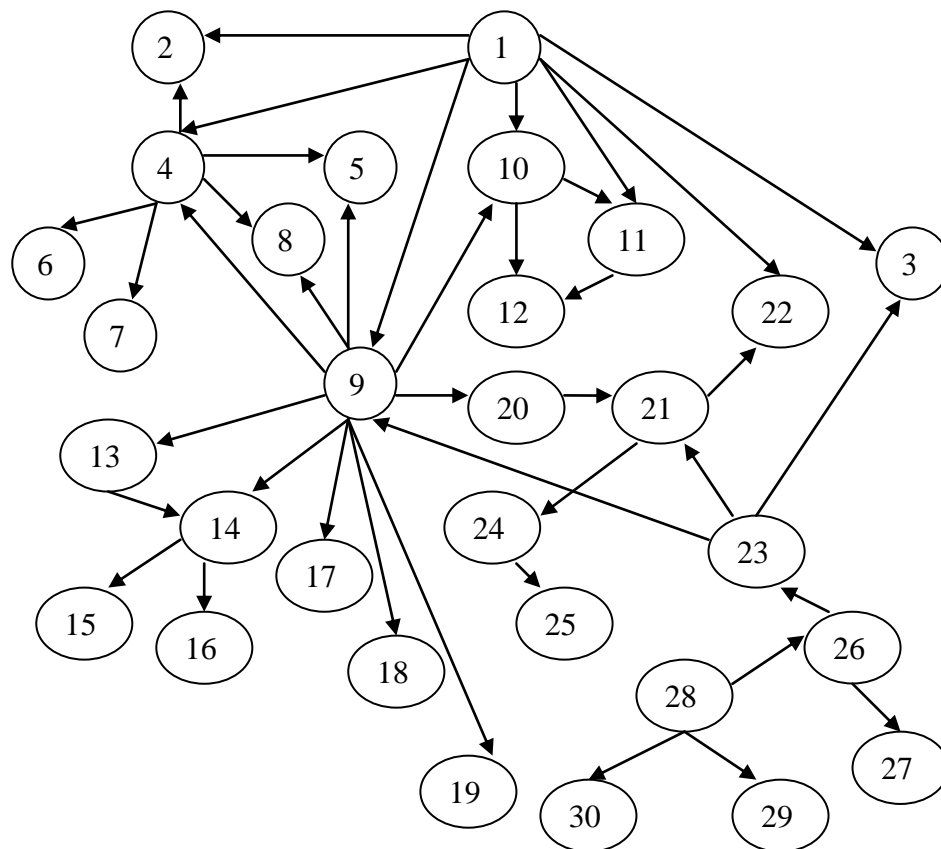


Рис. 2

1 – аренда оборудования; 2 – затраты на оборудование за прошлый месяц; 3 – перерасход средств на аренду оборудования; 4 – беспроводная связь; 5 – пейджинговая связь; 6 – затраты на беспроводную связь за прошедший месяц; 7 – перерасход средств на беспроводную связь; 8 – услуга голосовой почты; 9 – предоставляемый компанией пакет услуг; 10 – услуга электронного составления счетов; 11 – Интернет; 12 – уровень образования клиента; 13 – услуга режима ожидания; 14 – объём бесплатных услуг; 15 – перерасход бесплатных услуг; 16 – объём бесплатного трафика за прошлый месяц; 17 – услуга 3-way calling; 18 – услуга переадресации звонка; 19 – услуга автоопределения номера; 20 – услуга многоканальной связи; 21 – длительность пользования услугами компании; 22 – объём смешанного трафика за прошлый месяц; 23 – услуга пополнения счёта карточками; 24 – объём международного трафика; 25 – перерасход международного трафика; 26 – трудовой стаж; 27 – доход; 28 – возраст клиента; 29 – пенсионер; 30 – сколько лет проживает по адресу, указанному в анкете.

Приложение в медицине. ИАД используют для построения экспертных систем для постановки медицинских диагнозов. Такие системы построены главным образом на основе правил, описывающих сочетания различных симптомов различных заболеваний. С помощью вероятностного вывода в БС узнают не только, чем болен пациент, но и как нужно его лечить. Вероятностный вывод помогает выбирать средства медикаментозного воздействия, определять показания – противопоказания, ориентироваться в лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать исход назначенного курса лечения и прочее.

Пример 3. В медицинской базе накоплена информация о 1481 пациенте с сердечно-сосудистыми заболеваниями. На рис. 3 показана медицинская БС показывающая связь между заболеваниями пациента, хирургическим вмешательством и состоянием здоровья.

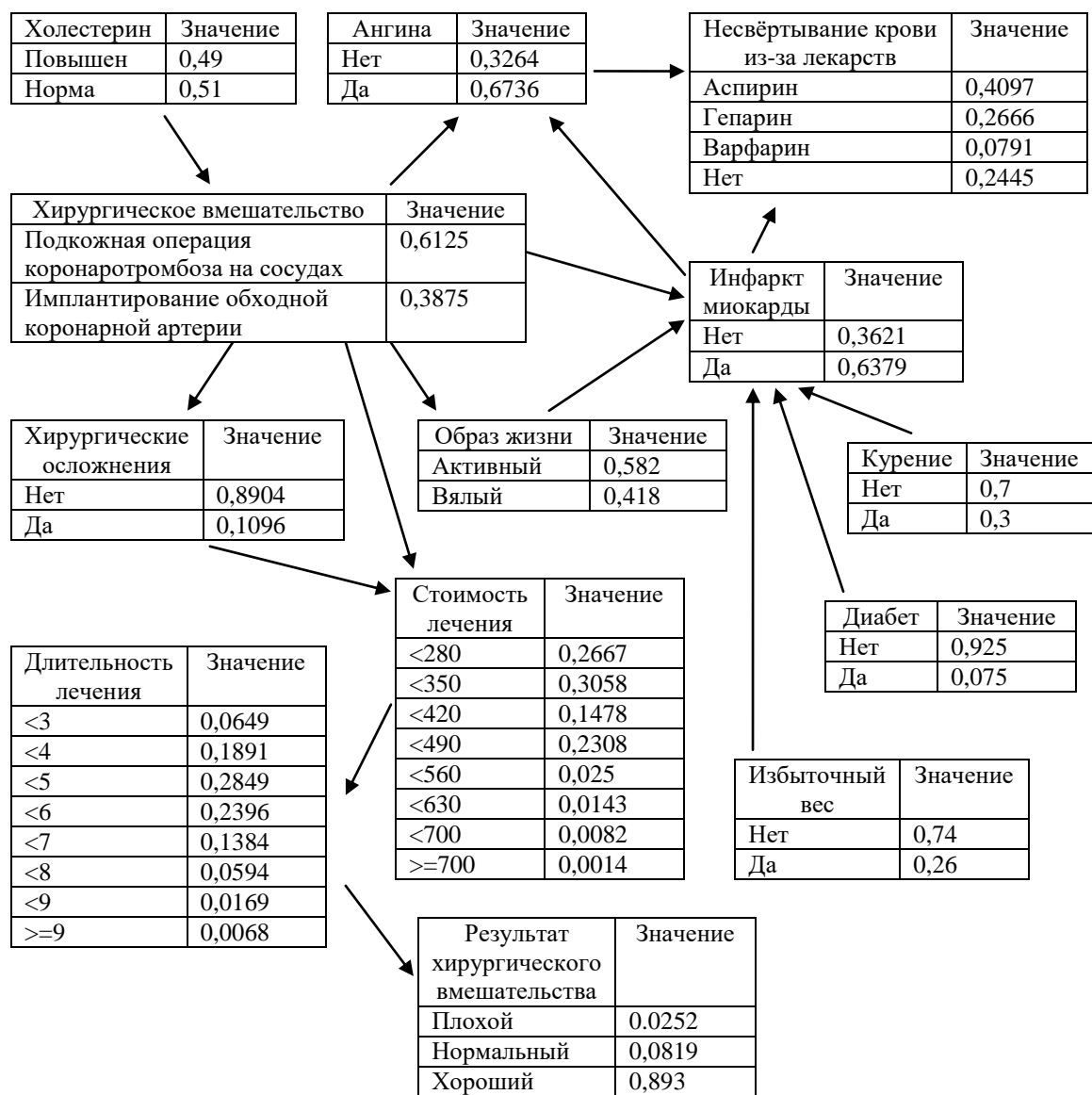


Рис. 3

Оценка потенциальных клиентов. Планируя выдачу кредита, банки требуют предоставления информации от потенциального клиента. Имея такую информацию, можно определить вероятность возврата кредита клиентом и, соответственно, можно принять решение о целесообразности выдачи кредита. Анализ клиентской базы позволяет выявить характерные особенности тех заявок, которые закончились невозвратами. Используя результаты данного анализа, менеджеры могут остановиться на более перспективных заявках клиентов.

Пример 4. За год работы банком было выдано 1194 кредита. Каждый клиент описывается 8 характеристиками. Рис. 4 иллюстрирует БС, построенную по этим данным. Она показывает связь между характеристиками клиента и вероятностью возврата кредита. Как видно из структуры сети, на решение банка о выдаче кредита клиенту главное влияние оказывает размер запрашиваемого кредита. Из этого можно сделать вывод, что банку нужно изменить метод сбора и состав информации о клиентах.

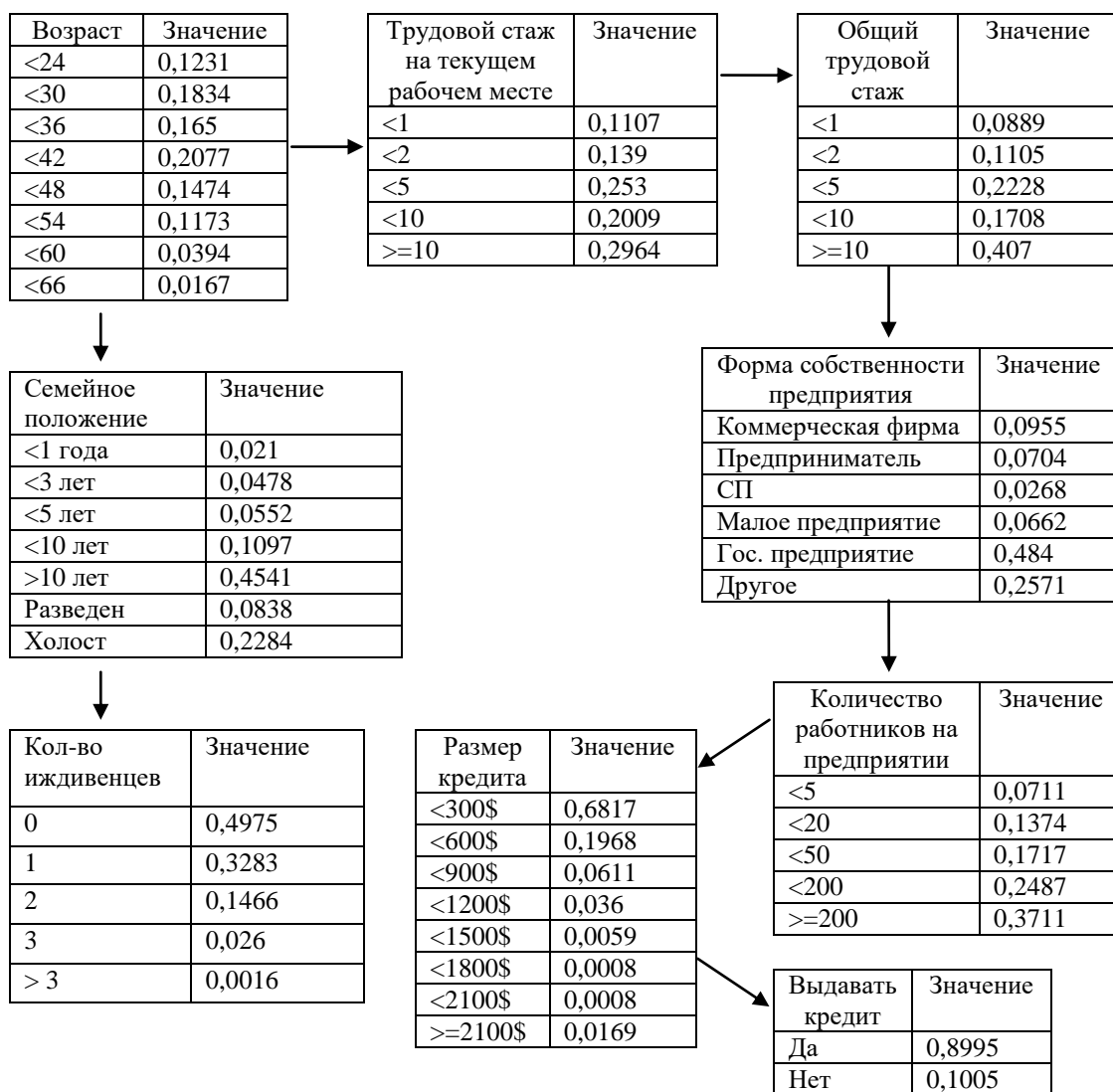


Рис. 4

Анализ и диагностика надежности и отказоустойчивости технических систем.

Выполнение профилактического ремонта по мере необходимости обходится дешевле, чем устранение последствий технического сбоя или дублирование компонентов оборудования. Выполняя анализ состояния отдельных компонентов технической системы можно оценивать вероятность выхода системы или её частей из рабочего состояния. Используя результаты такого анализа, технический персонал может выполнять ремонт не всей системы, а наиболее проблемных компонентов.

Пример 5. На рис. 5 представлена БС показывающая связь между разными компонентами автомобиля и вероятностью того, что автомобиль заведётся.

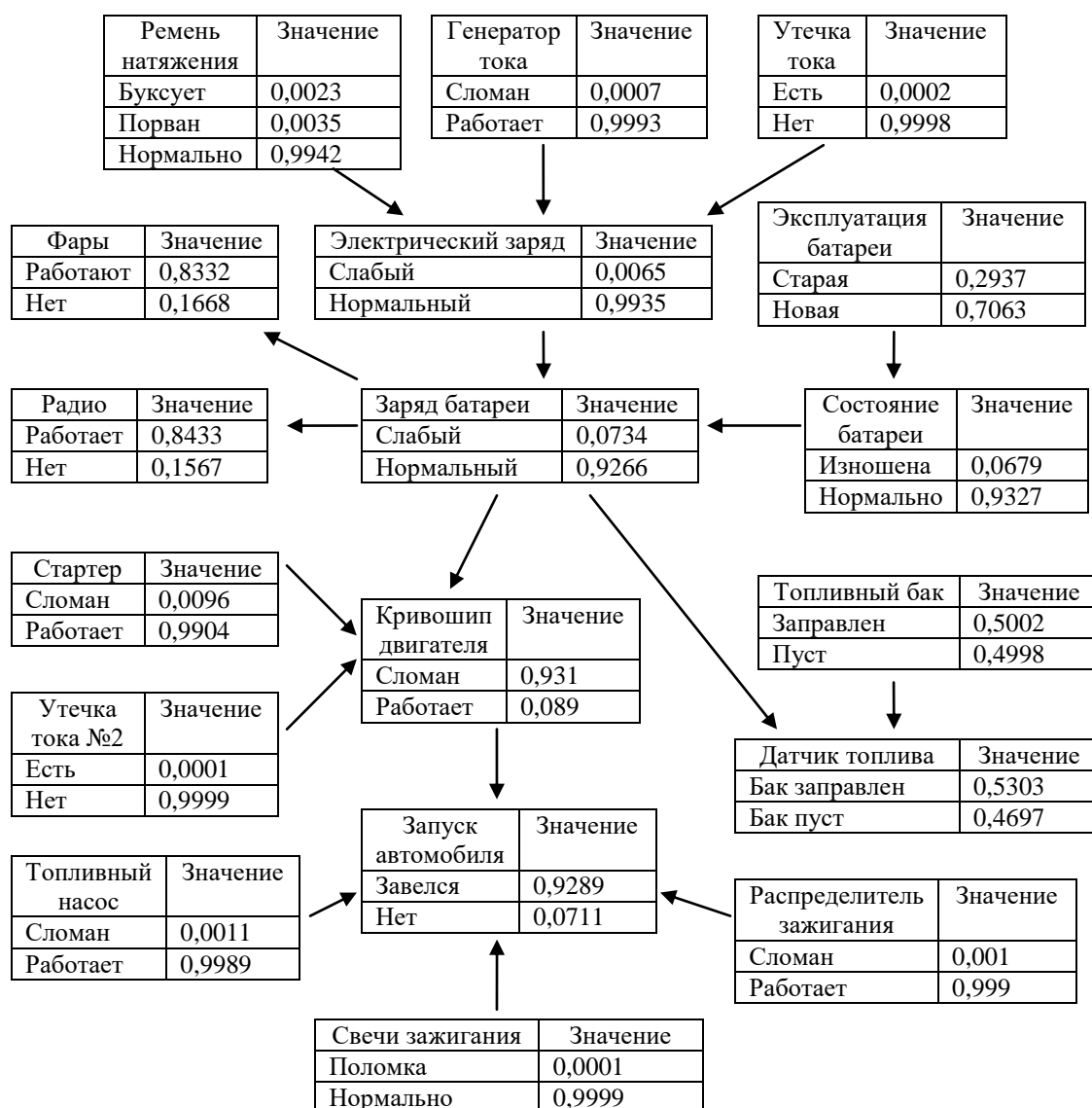


Рис. 5

Очевидно, что перечисленные виды задач актуальны практически для всех отраслей и направлений человеческой деятельности – банковского дела и страхования (выявление злоупотреблений с кредитными карточками, оценки кредитных рисков, оценки залладных, выявления профилей пользователей, оценки эффективности региональных отделений,

вероятности подачи заявки на выплату страховки), финансовых рынков (прогнозирование, анализ портфелей, моделирование индексов), производства (прогнозирование спроса, контроль качества, оценка дизайна продукции), торговли, медицине, анализ и диагностика надежности технических систем и прочее.

Экспериментальные результаты

Для построения всех пяти, выше перечисленных практических приложений БС, был использован алгоритм эвристического метода построения БС. Данные для обучения БС взяты из Интернета: <http://ftp.njfiw.gov.cn/ww/SPSS11.5/Tutorial/SAMPLE~1>. В таблице показаны вычислительные характеристики каждого проведенного вычислительного эксперимента.

Таблица

Пример	Кол-во вершин БС	Кол-во итераций алгоритма	Количество проанализированных моделей	Затраченное время на вычисление	Размер обучающей выборки
1	23	252	3329	5 мин. 4 сек.	6400
2	30	434	5065	6 мин. 44 сек.	1000
3	14	90	686	1 мин. 11 сек.	1481
4	9	35	295	5 сек.	1194
5	18	152	600	28 сек.	10000

Как видно из таблицы, скорость построения оптимальной структуры БС в первую очередь зависит от количества вершин сети.

Выводы

Моделирование процессов различной природы и сложности при помощи БС является одним из перспективных современных направлений в области ИАД. В статье предложены: (1) итерационный алгоритм эвристического метода построения БС, позволяющий значительно уменьшить вычислительную сложность обучения БС; (2) метод формирования вероятностного вывода на основе БС с использованием обучающих данных. Также приведены экспериментальные данные, полученные при моделировании различных процессов сетями Байеса, из которых видно, что вычислительная сложность алгоритма напрямую зависит от количества переменных принимающих участие в анализе.

В дальнейшем планируется разработка более совершенных методов вероятностного вывода для БС с непрерывными вершинами, т.е. переменными, которые подчиняются стандартным законам распределения, а также более интересного случая БС с неполными обучающими данными.

Список литературы

1. Robinson R.W. Counting unlabeled acyclic digraphs / Proceeding of fifth Australian conference on combinatorial mathematics, Melbourne, Australia, 1976. – P. 28-43.
2. Leray P., Francois O. BNT structure learning package: documentation and experiments / Technical report, laboratory PSI-INSa Rouen-FRE CNRS 2645, November 2004. – 27 p.
3. Lokeswarappa K.G. Junction trees: motivation / Seminar CSE 714 on advanced topics in machine learning, March 2005. – 57 p.
4. Dechter R. Bucket elimination: a unifying framework for reasoning // ACM Press, Vol. 28, article № 61, December 1996. – P. 1-51.
5. Huang C., Darwiche A. Inference in belief networks: g procedural guide // International journal of approximate reasoning. – 1994. – 11. – P. 1-45.
6. Lepar V., Shenoy P. A Comparison of Lauritzen-Spiegelhalter, Hugin, and Shenoy-Shafer architectures for computing marginals of probability distributions // Uncertainty in artificial intelligence. – 1999. – 14. – P. 328-337.
7. Murphy K.P. Dynamic Bayesian networks: representation, inference and learning / A PhD dissertation, University of California, Berkeley, 2002. – 225 p.
8. Бидюк П.И., Терентьев А.Н., Гасанов А.С. Построение и методы обучения Байесовских сетей // Кибернетика и системный анализ. – 2005. – 4. – С. 133-147.
9. Терентьев А.Н., Бидюк П.И. Методы построения Байесовских сетей // Адаптивные системы автоматического управления (межведомственный научно-технический сборник). – 2005. – 8. – С. 130 – 141.
10. Терентьев А.Н., Бидюк П.И., Эвристический метод построения Байесовских сетей // Математические машины и системы. – 2006. – 3. – С. 12-23.

Ключевые слова: Байесовская сеть, описание минимальной длины (ОМД), обоюдная информация, эвристический метод обучения, вероятностный вывод.

Ключові слова: Байєсова мережа, описання мінімальної довжини (ОМД), взаємна інформація, евристичний метод навчання, ймовірнісний висновок.

Key words: Bayesian networks, minimum description length (MDL), mutual information, heuristic method of learning, probabilistic inference.

УДК 62-50

Аннотация

Терентьев А.Н., Бидюк П.И., Коршевнюк Л.А.

Байесовская сеть – инструмент интеллектуального анализа данных

Байесовские сети (БС) являются серьёзным востребованным инструментом для выполнения интеллектуального анализа данных различной природы. Однако существующие методы построения БС и вероятностного вывода имеют вычислительную сложность. Для построения БС предлагается использовать эвристический метод построения БС, а для выполнения вероятностного вывода – алгоритм вероятностного вывода в БС на основе обучающих данных. В статье приведены результаты экспериментов и примеры использования БС.

Рис. 5.: Табл. 1.: Библиогр. 10 названий.

УДК 62-50

Анотація

Терентьєв А.Н., Бідюк П.И., Коршевнюк Л.О.

Байєсова мережа – інструмент інтелектуального аналізу даних

Байєсові мережі (БМ) – потужний інструмент для інтелектуального аналізу даних різної природи. Однак існуючі методи побудови БМ і формування ймовірнісного висновку мають високу обчислювальну складність. Для побудови БМ пропонується застосовувати евристичний метод побудови БМ, а для формування ймовірнісного висновку – алгоритм формування ймовірнісного виводу в БС на основі навчальних даних. Експериментальні результати та приклади застосування БМ наведені в статті.

Іл. 5.: Табл. 1.: Бібліогр. 10.

UDC 62-50

Abstract

Terentyev O.M., Bidyuk P.I., Korchevnyuk L.O.

Bayesian networks – instrument of data mining

Bayesian networks represent a useful and seriously claimed instrument for implementation in data-mining systems of various applications. Existing methods of construction BN and probabilistic inference from training data have high computation complexity. For construction BN are propose heuristic method of construction BN, and for implementation of probabilistic inference are propose an algorithm of probabilistic inference in BN from a database. Experimental results and examples for using BN are presented.

Figs. 5.: Tabl. 1.: Refs. 10.