МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МИНИСТЕРСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ ПО АТОМНОЙ ЭНЕРГИИ

МИНИСТЕРСТВО ПРОМЫШЛЕННОСТИ, НАУКИ И ТЕХНОЛОГИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ

РОССИЙСКАЯ АССОЦИАЦИЯ НЕЙРОИНФОРМАТИКИ

МОСКОВСКИЙ ИНЖЕНЕРНО-ФИЗИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

НАУЧНАЯ СЕССИЯ МИФИ–2003

НЕЙРОИНФОРМАТИКА–2003

**V ВСЕРОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ**

ЛЕКЦИИ
ПО НЕЙРОИНФОРМАТИКЕ

Часть 1

По материалам Школы-семинара
«Современные проблемы нейроинформатики»

Москва 2003

УДК 004.032.26 (06)
ББК 32.818я5
М82

НАУЧНАЯ СЕССИЯ МИФИ–2003. V ВСЕРОССИЙСКАЯ НАУЧНО-ТЕХНИЧЕСКАЯ КОНФЕРЕНЦИЯ «НЕЙРОИНФОРМАТИКА–2003»: ЛЕКЦИИ ПО НЕЙРОИНФОРМАТИКЕ. Часть 1. – М.: МИФИ, 2003. – 188 с.

В книге публикуются тексты лекций, прочитанных на Школе-семинаре «Современные проблемы нейроинформатики», проходившей 29–31 января 2003 года в МИФИ в рамках V Всероссийской конференции «Нейроинформатика–2003».

Материалы лекций связаны с рядом проблем, актуальных для современного этапа развития нейроинформатики, включая ее взаимодействие с другими научно-техническими областями.

Ответственный редактор
Ю. В. Тюменцев, кандидат технических наук

ISBN 5–7262–0471–9

Содержание

С. А. ТЕРЕХОВ
Снежинский физико-технический институт, г. Снежинск;
ООО НейрОК, г. Москва,
E-mail: alife@narod.ru

ВВЕДЕНИЕ В БАЙЕСОВЫ СЕТИ

Аннотация

Байесовы сети представляют собой графовые модели вероятностных и причинно-следственных отношений между переменными в статистическом информационном моделировании. В байесовых сетях могут органически сочетаться эмпирические частоты появления различных значений переменных, субъективные оценки «ожиданий» и теоретические представления о математических вероятностях тех или иных следствий из априорной информации. Это является важным практическим преимуществом и отличает байесовы сети от других методик информационного моделирования. В обзорной лекции изложено краткое введение в методы вычисления вероятностей и вывода статистических суждений в байесовых сетях.

S. A. TEREKHOFF
Snezhinsk Institute of Physics and Technology (SFTI), Snezhinsk;
NeurOK LLC, Moscow,
E-mail: alife@narod.ru

BAYESIAN NETWORKS PRIMER

Abstract

Bayesian networks represent probabilistic graph models of casual relations between variables in statistical information modelling. Bayesian networks consolidate empirical frequencies of events, subjective beliefs and theoretical representations of mathematical probabilities. It is the important practical advantage which distinguishes Bayesian networks from other techniques of information modelling. This survey lecture presents a brief introduction to calculation of probabilities and statistical inference methods in Bayesian networks.

УДК 004.032.26 (06) Нейронные сети

Вероятностное представление знаний в машине

Наблюдаемые события редко могут быть описаны как прямые следствия строго детерминированных причин. На практике широко применяется вероятностное описание явлений. Обоснований тому несколько: и наличие неустранимых погрешностей в процессе экспериментирования и наблюдений, и невозможность полного описания структурных сложностей изучаемой системы [26], и неопределенности вследствие конечности объема наблюдений.

На пути вероятностного моделирования встречаются определенные сложности, которые (если отвлечься от чисто теоретических проблем) можно условно разделить на две группы:

- технические (вычислительная сложность, «комбинаторные взрывы» и т. п.);
- идейные (наличие неопределенности, сложности при постановке задачи в терминах вероятностей, недостаточность статистического материала).

Для иллюстрации еще одной из «идейных» сложностей рассмотрим простой пример из области вероятностного прогнозирования [5]. Требуется оценить вероятность положительного исхода в каждой из трех ситуаций:

- Знатная леди утверждает, что она может отличить на вкус, был ли чай налит в сливки или наоборот — сливки в чай. Ей удалось это проделать 10 раз в течение бала.
- Азартный игрок утверждает, что он может предсказать, орлом или решкой выпадет монета (которую вы ему дадите). Он смог выиграть такое пари уже 10 раз за этот вечер, ни разу не проиграв!
- Эксперт в классической музыке заявляет, что он в состоянии различить творения Гайдна и Моцарта лишь по одной странице партитуры. Он уверенно проделал это 10 раз в музыкальной библиотеке.

Удивительная особенность — во всех трех случаях мы формально имеем одинаковые экспериментальные свидетельства в пользу высказанных утверждений — в каждом случае они достоверно подтверждены 10 раз. Однако мы с восхищением и удивлением отнесемся к способностям леди, весьма скептически воспримем заявления бравого игрока, и совершенно естественно согласимся с доводами музыкального эксперта. Наши субъективные оценки вероятности этих трех ситуаций весьма отличаются. И, несмотря на то, что мы имеем дело с повторяющимися событиями, весьма

непросто совместить их с классическими положениями теории вероятностей.

Особенно затруднительно получить формулировку, понятную вычислительной машине.

Другая сторона идейных трудностей возникает при практической необходимости вероятностного прогнозирования событий, к которым не вполне применимы классические представления о статистической повторяемости. Представим себе серию экспериментов с бросанием кубика, сделанного из сахара, на влажную поверхность стола. Вероятности исходов последующих испытаний зависят от относительной частоты исходов предыдущих испытаний, при этом исследуемая система каждый раз необратимо изменяется в результате каждого эксперимента. Этим свойством обладают многие биологические и социальные системы, что делает их вероятностное моделирование классическими методами крайне проблематичным.

Часть из указанных проблем решается в вероятностных *байесовых сетях*, которые представляют собой графовые модели причинно-следственных отношений между случайными переменными. В байесовых сетях могут органически сочетаться эмпирические частоты появления различных значений переменных, субъективные оценки «ожиданий» и теоретические представления о математических вероятностях тех или иных следствий из априорной информации. Это является важным практическим преимуществом и отличает байесовы сети от других методик информационного моделирования.

Байесовы сети широко применяются в таких областях, как медицина, стратегическое планирование, финансы и экономика.

Изложение свойств байесовых сетей, которые являются основным предметом этой лекции, будет построено следующим образом. Вначале рассматривается общий байесов подход к моделированию процесса вывода суждений в условиях неопределенности. Далее излагаются алгоритмы вычисления вероятностей сложных событий в байесовых сетях. Затем обсуждаются подходы к обучению параметров таких сетей. В завершение лекции предлагается эффективная практическая методика эмпирического вывода на основе вероятностных деревьев, и рассматриваются приложения байесовых методик. В приложении приведен краткий обзор ресурсов Интернет по данной тематике.

Неопределенность и неполнота информации

Экспертные системы и формальная логика

Попробуем проследить за способом работы эксперта в некоторой определенной области. Примерами экспертов являются врач, проводящий обследование, финансист, изучающий условия предоставления ссуды, либо пилот, управляющий самолетом.

Действия эксперта могут условно быть представлены в виде повторяющейся последовательности из трех этапов:

- получение информации о состоянии окружающего мира;
- принятие решения относительно выбора некоторых действий, по поводу которых у эксперта имеются определенные ожидания последствий;
- приобретение опыта путем сопоставления результатов действий и ожиданий и возврат к первому этапу.

Приобретенные новый опыт и информация о мире позволяют эксперту сообразно действовать в будущем.

Попытки компьютерного моделирования действий эксперта привели в конце 60-х годов к появлению экспертных систем (ЭС) [6], которые чаще всего основывались на продукционных правилах типа «ЕСЛИ условие, ТО факт или действие». Будущее подобных систем связывалось при этом с заменой экспертов их моделями. Однако после первых успехов обнажились проблемы, и первой среди них — серьезные затруднения при попытках работы с нечеткой, недоопределенной информацией.

Следующие поколения ЭС претерпели кардинальные изменения:

- вместо моделирования эксперта моделируется предметная область;
- вместо попыток учета неопределенности в правилах — использование классической теории вероятностей и теории принятия решений;
- вместо попыток замены эксперта — оказание ему помощи.

В конце 80-х годов были предложены обобщения ЭС в виде байесовых [1] сетей [21] и была показана практическая возможность вычислений вероятностных выводов даже для сетей больших размеров.

[1] По-видимому, первые работы по использованию графовых моделей в статистическом моделировании относятся еще к 20-м годам прошлого века. (*Wright S.* Correlation and causation // *Journal of Agricultural Research*, **20**, 557, 1921).

Вернемся к трехэтапному описанию профессиональных действий эксперта. Сейчас нас будет интересовать вопрос, как наблюдения эксперта, т. е. получение им информации о внешнем мире, изменяют его ожидания по поводу ненаблюдаемых событий?

Особенности вывода суждений в условиях неопределенности

Суть приобретаемого знания в условиях неопределенности состоит в понимании, влияет ли полученная информация на наши ожидания относительно других событий. Основная причина трудностей при использовании систем, основанных на правилах, состоит в учете «сторонних», «косвенных» последствий наблюдаемых событий. Проиллюстрируем это на уже успевшем стать классическим примере [15].

Шерлок Холмс вышел из дома утром и заметил, что трава вокруг влажная. Он рассудил [2]: «Я думаю, что ночью был дождь. Следовательно, трава возле дома моего соседа, доктора Ватсона, вероятно, также влажная». Таким образом, информация о состоянии травы у дома Холмса *повлияла на его ожидания* относительно влажности травы у дома Ватсона. Но предположим, что Холмс проверил состояние сборника дождевой воды и обнаружил, что тот — сухой. В результате Холмс вынужден изменить ход своих рассуждений, и состояние травы возле его дома *перестает* влиять на ожидания по поводу травы у соседа.

Теперь рассмотрим две возможные причины, почему трава у дома Холмса оказалась влажной. Помимо дождя, Холмс мог просто забыть выключить поливальную установку накануне. Допустим, на следующее утро Холмс снова обнаруживает, что трава влажная. Это повышает его субъективные вероятности и для прошедшего дождя, и по поводу забытой дождевальной установки. Затем Холмс обнаруживает, что трава у дома Ватсона также влажная и заключает, что ночью был дождь.

Следующий шаг рассуждений практически невозможно воспроизвести в системах, основанных на правилах, однако он абсолютно естественен для человека: *влажность травы у дома Холмса объясняется дождем, и следовательно нет оснований продолжать ожидать, что была забыта включенной поливальная машина. Следовательно, возросшая, было, субъек-*

[2] Мы сознательно обужаем мир возможных причин и способов рассуждений великого сыщика до элементарного их набора, что не мешает, однако, понять, почему и они весьма трудно программируются в машине.

тивная вероятность относительно забытой поливальной машины уменьшается до (практически) исходного значения, имевшего место до выхода Холмса из дома. Такой способ рассуждения можно назвать «попутное объяснение», «контекстное объяснение» или «редукция причины» (explaining away).

Важная особенность «попутного объяснения» состоит в изменении отношений зависимости между событиями по мере поступления информации. До выхода из дома Холмса факты дождя и работы поливальной установки были независимы. После получения информации о траве у дома они стали зависимыми. Далее, когда появилась информации о влажности травы у дома Ватсона, состояние зависимости вновь изменилось.

Эту ситуацию удобно описать при помощи графа, узлы которого представляют события (или переменные), а пара узлов (A, B) связывается направленным ребром, если информация об A может служить причиной для B. В этом случае узел A будет родителем для B, который, в свою очередь, называется узлом-потомком по отношению к A.

История с травой у Холмса и Ватсона представлена на рис. 1.

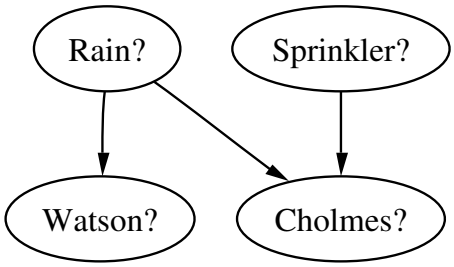

Рис. 1. Граф рассуждений Шерлока Холмса

Граф на рис. 1 может быть отнесен к семейству *байесовых сетей*. В данном примере переменные в узлах могут принимать только булевы значения 1 или 0 (да/нет). В общем случае переменная может принимать одно из набора [3] взаимоисключающих состояний. Из графа на рис. 1 можно сделать несколько полезных выводов о зависимости и независимости переменных.

[3] В традиционной постановке байесовы сети не предназначены для оперирования с непрерывным набором состояний (например, с действительным числом на заданном отрезке). Для представления действительных чисел в некоторых приложениях можно провести разбиение отрезка на сегменты и рассматривать дискретный набор их центров.

Например, если известно, что ночью не было дождя, то информация о состоянии травы у дома Ватсона не оказывает влияния на ожидания по поводу состояния травы у дома Холмса.

В середине 80-х годов были подробно проанализированы способы, которыми влияние информации распространяется между переменными в байесовой сети [21]. Будем считать, что две переменные разделены, если новые сведения о значении одной из них не оказывают влияния на ожидания по поводу другой. Если состояние переменной известно, мы будем называть такую переменную конкретизированной. В байесовой сети возможны три типа отношений между переменными: *последовательные* соединения (рис. 2a), *дивергентные* соединения (рис. 2b), *конвергентные* соединения (рис. 2c).

Ситуация на рис. 2c требует, по-видимому, дополнительных пояснений — как возникает зависимость между предками конвергентного узла, когда становится известным значение потомка [4]. Для простоты рассмотрим пример, когда узел A имеет всего двух предков — B и C. Пусть эти две переменные отвечают за выпадение орла и решки при независимом бросании двух разных монет, а переменная A — логический индикатор, который «загорается», когда обе монеты оказались в одинаковом состоянии (например, обе — решки). Теперь легко понять, что если значение индикаторной переменной стало *известным*, то значения B и C стали *зависимыми* — знание одного из них полностью определяет оставшееся.

Общее свойство (условной) независимости переменных — узлов в байесовой сети получило название d-разделения (d-separation).

Определение (d-разделимость). Две переменные A и B в байесовой сети являются d-разделенными, если на каждом пути, соединяющем эти две вершины на графе, найдется промежуточная переменная V, такая что:

- соединение с V последовательное или дивергентное, и значение V известно, либо
- соединение конвергентное, и нет свидетельств ни о значении V, ни о каждом из ее потомков.

Так, в сети задачи Шерлока Холмса (рис. 1) переменные «Полив?» и «Трава у дома Ватсона?» являются d-разделенными. Граф содержит на пути между этими переменными конвергентное соединение с переменной «Трава у дома Холмса?», причем ее значение достоверно известно (трава — влажная).

[4] Зависимости, возникшие таким способом, называют *индуцированными*.

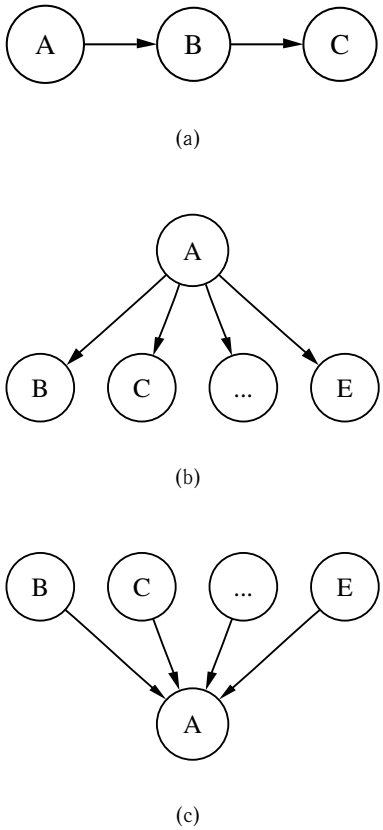

Рис. 2. Три типа отношений между переменными

(a) *Последовательное соединение*. Влияние информации может распространяться от A к C и обратно, пока значение B не конкретизировано. **(b)** *Дивергентное соединение*. Влияние может распространяться между потомками узла A, пока его значение не конкретизировано. **(c)** *Конвергентное соединение*. Если об A ничего не известно, кроме того, что может быть выведено из информации о его предках $B, C, \ldots, E$, то эти переменные предки являются разделенными. При уточнении A открывается канал взаимного влияния между его предками.

Свойство d-разделимости соответствует особенностям логики эксперта-человека, поэтому крайне желательно, чтобы в рассуждениях машин относительно двух d-разделенных переменных новая информация об одной из них не изменяла степень детерминированности второй переменной. Формально, для переменных A и C, независимых при условии B, имеет место соотношение $P(A \mid B) = P(A \mid B, C)$.

Отметим, что интуитивное восприятие условной зависимости и независимости иногда, даже в простых случаях, оказывается затрудненным, так как сложно из всех исходов событий мысленно выделить только те события, в которых значение обусловливающей переменной определено, и далее в рассуждения оперировать только ими.

Вот простой пример, поясняющий эту трудность: в некотором сообществе мужчины среднего возраста и молодые женщины оказались материально более обеспеченными, чем остальные люди. Тогда *при условии фиксированного повышенного уровня обеспеченности* пол и возраст человека оказываются условно *зависимыми* друг от друга!

Еще один классический пример, связанный с особенностями условных вероятностей. Рассмотрим некоторый колледж, охотно принимающий на обучение сообразительных и спортивных молодых людей (и тех, кто обладает обоими замечательными качествами!). Разумно считать, что среди *всех* молодых людей студенческого возраста спортивные и интеллектуальные показатели независимы. Теперь если вернуться к множеству зачисленных в колледж, то легко видеть, что высокая сообразительность эффективно *понижает* вероятность спортивности и наоборот, так как каждого из этих свойств *по-отдельности* достаточно для приема в колледж. Таким образом, спортивность и умственные показатели *оказались* зависимыми при условии обучения в колледже.

Исчисление вероятностей и байесовы сети

Вероятности прогнозируемых значений отдельных переменных

Рассмотрим вначале вопрос о вычислении вероятностей интересующих нас значений переменных в обученной байесовой сети при условии, что в нашем распоряжении имеется некоторая информация о значениях (части) других переменных.

Это — задача статистического вывода суждений. В простейших случаях

она может быть решена точно с использованием соотношений для условных вероятностей. Понятие условной вероятности $P(A \mid B) = x$ составляет основу байесова подхода к анализу неопределенностей. Приведенная формула означает «при условии, что произошло B (и всего остального, что не имеет отношения к A), вероятность возникновения A равна x». Совместная вероятность наступления событий A и B дается формулой полной вероятности:

$$P(A, B) = P(A \mid B) \cdot P(B)\,.$$

Если в нашем распоряжении имеется информация о зависимых переменных (*следствиях*), а суть исследования состоит в определении сравнительных вероятностей исходных переменных (*причин*), то на помощь приходит теорема Байеса.

Пусть имеется условная вероятность $P(A \mid B)$ наступления некоторого события A при условии, что наступило событие B. Теорема Байеса дает решение для *обратной* задачи — какова вероятность наступления более раннего события B, если известно что более позднее событие A наступило. Более точно, пусть $A_1, \ldots, A_n$ — набор (полная группа) несовместных взаимоисключающих событий (или альтернативных гипотез). Тогда *апостериорная* вероятность $P(A_j \mid B)$ каждого их событий A_j при условии, что произошло событие B, выражается через априорную вероятность $P(A_j)$:

$$P(A_j \mid B) = \frac{P(A_j) \cdot P(B \mid A_j)}{P(B) = \sum_{j=1}^{n} P(A_j) \cdot P(B \mid A_j)}\,.$$

Обратная вероятность $P(B \mid A_j)$ называется *правдоподобием* (likelihood), а знаменатель P(B) в формуле Байеса — *свидетельством* (evidence) [5]. Совместная вероятность является наиболее полным статистическим описанием наблюдаемых данных. Совместное распределение представляется функцией многих переменных — по числу исследуемых переменных в задаче. В общем случае это описание требует задания вероятностей всех

[5] Термин “evidence” не получил общеупотребительного аналога в отечественной литературе, и его чаще называют просто «знаменателем в формуле Байеса». Важность evidence проявляется при сравнительном теоретическом анализе различных моделей, статистически объясняющих наблюдаемые данные. Более предпочтительными являются модели, имеющие наибольшее значение evidence. К сожалению, на практике вычисление evidence часто затруднено, так как требует суммирования по всем реализациям параметров моделей. Популярным способом приближенного оценивания evidence являются специализированные методы Монте-Карло.

допустимых конфигураций значений всех переменных, что мало применимо даже в случае нескольких десятков булевых переменных. В байесовых сетях, в условиях, когда имеется дополнительная информация о степени зависимости или независимости признаков, эта функция факторизуется на функции меньшего числа переменных:

$$P(A_1, \ldots, A_n) = \prod_j P\big[A_j \mid pa(A_j)\big] .$$

Здесь $pa(A_j)$ — состояния всех переменных-предков для переменной A_j. Это выражение носит название *цепного правила* для полной вероятности.

Важно, что обусловливание происходит всей совокупностью переменных-предков A_j — в противном случае будет потеряна информация об эффектах совместного влияния этих переменных.

Таким образом, байесова сеть состоит из следующих понятий и компонент:

- множество случайных переменных и направленных связей между переменными;
- каждая переменная может принимать одно из конечного множества взаимоисключающих значений;
- переменные вместе со связями образуют ориентированный граф без циклов;
- каждой переменной-потомку A с переменными-предками $B_1, \ldots, B_n$ приписывается таблица условных вероятностей $P(A \mid B_1, \ldots, B_n)$.

Если переменная A не содержит предков на графе, то вместо условных вероятностей (автоматически) используются безусловные вероятности $P(A)$. Требование отсутствия (ориентированных) петель является существенным — для графов с петлями в цепочках условных вероятностей в общем случае нет корректной схемы проведения вычислений — вследствие бесконечной рекурсии.

На практике нам необходимы распределения интересующих нас переменных, взятые по отдельности. Они могут быть получены из соотношения для полной вероятности при помощи *маргинализации* — суммирования по реализациям всех переменных, кроме выбранных.

Приведем пример точных вычислений в простой байесовой сети, моделирующей задачу Шерлока Холмса. Напомним обозначения и смысл переменных в сети (рис. 1): R — был ли дождь, S — включена ли поливальная

установка, C — влажная ли трава у дома Холмса, и W — влажная ли трава у дома Ватсона.

Все четыре переменные принимают булевы значения 0 — ложь, (f) или 1 — истина (t). Совместная вероятность $P(R,S,C,W)$, таким образом, дается совокупной таблицей из 16 чисел. Таблица вероятностей нормирована, так что

$$\sum_{R=\{t,f\},\ldots,W=\{t,f\}} P(R,S,C,W) = 1\,.$$

Зная совместное распределение, легко найти любые интересующие нас условные и частичные распределения. Например, вероятность того, что ночью не было дождя при условии, что трава у дома Ватсона — влажная, дается простым вычислением:

$$P(R=f \mid W=t) = \frac{P(R=f, W=t)}{P(W=t)} =$$

$$= \frac{\sum_{S=\{t,f\},C=\{t,f\}} P(R=f,S,C,W=t)}{\sum_{R=\{t,f\},S=\{t,f\},C=\{t,f\}} P(R,S,C,W=t)}\,.$$

Из теоремы об умножении вероятностей [7, с. 45] полная вероятность представляется цепочкой условных вероятностей:

$$P(R,S,C,W) = P(R)\cdot P(S \mid R)\cdot P(C \mid R,S)\cdot P(W \mid R,S,C)\,.$$

В описанной ранее байесовой сети ориентированные ребра графа отражают суть те обусловливания вероятностей, которые реально имеют место в задаче. Поэтому формула для полной вероятности существенно упрощается:

$$P(R,S,C,W) = P(R)\cdot P(S)\cdot P(C \mid R,S)\cdot P(W \mid R)\,.$$

Порядок следования переменных в соотношении для полной вероятности, вообще говоря, может быть любым. Однако на практике целесообразно выбирать такой порядок, при котором условные вероятности максимально редуцируются. Это происходит, если начинать с переменных-«причин», постепенно переходя к «следствиям». При этом полезно представлять себе некоторую «историю», согласно которой причины влияют на следствия [6].

[6] Существуют методики частичной автоматизации выбора порядка переменных и синтеза топологии байесовой сети, основанные на большом объеме экспериментальных данных — примеров связей между переменными в данной задаче.

ТАБЛИЦА 1. Априорные условные вероятности в байесовой сети

$P(R=t)$	$P(R=f)$
0.7	0.3

$P(S=t)$	$P(S=f)$
0.2	0.8

R	$P(W=t \mid R)$	$P(W=f \mid R)$
t	0.8	0.2
f	0.1	0.9

R	S	$P(R=t)$	$P(R=f)$
t	t	0.9	0.1
t	f	0.8	0.2
f	t	0.7	0.3
f	f	0.1	0.9

В этом смысле, исключая связь между двумя переменными в сети, мы удаляем зависимость от этой переменной в конкретном выражении для условной вероятности, делая переменные условно независимыми [7].

В байесовой сети для описания полной вероятности требуется указание таблиц локальных условных вероятностей для всех дочерних переменных в зависимости от комбинации значений их предков на графе.

Эта таблица (по случайности!) также содержит 16 чисел, однако вследствие разбиения задачи на мелкие подзадачи с малым числом переменных в каждой, нагрузка на эксперта по их оцениванию значительно снижена.

Перейдем к вычислениям. Найдем сначала полные вероятности двух событий: трава у дома Холмса оказалась влажной, и у дома Ватсона наблюдается то же самое.

$$P(C=t) = \sum_{\substack{R=\{t,f\}\\S=\{t,f\}}} P(R)\cdot P(S)\cdot P(C=t \mid R,S) =$$

$$= 0.3\cdot 0.2\cdot 0.9 + 0.3\cdot 0.8\cdot 0.8 + 0.7\cdot 0.2\cdot 0.7 + 0.7\cdot 0.8\cdot 0.1 = 0.4\,.$$

[7] В литературе часто встречается утверждение «отсутствие ребра в байесовой сети означает независимость между соответствующими переменными». Понимать его следует как условную независимость этих переменных *при отсутствии прочей информации*. Однако, как отмечено в тексте, эти две переменных могут оказаться зависимыми, если они являются предками конвергентного узла, *при наличии информации* о значении в этом узле.

Аналогично,

$$P(W = t) = \sum_{R=\{t,f\}} P(R) \cdot P(W = t \mid R) = 0.31\,.$$

Теперь, если мы достоверно знаем, что ночью был дождь, то эта информация изменит (увеличит) вероятности наблюдения влаги на траве:

$$P(C = t \mid R = t) = \sum_{S=\{t,f\}} P(S) \cdot P(C = t \mid R = t, S) = 0.82\,,$$
$$P(W = t \mid R = t) = 0.8\,.$$

С другой стороны, пусть Холмсу известно, что трава у его дома влажная. Каковы вероятности того, что был дождь, и что дело — в поливальной установке? Вычисления дают:

$$P(R = t \mid C = t) = \frac{P(R = t, C = t)}{P = (C = t)} = \frac{0.054 + 0.192}{0.4} = 0.615\,.$$

$$P(S = t \mid C = t) = \frac{P(S = t, C = t)}{P = (C = t)} = \frac{0.054 + 0.098}{0.4} = 0.38\,.$$

Числитель этой формулы получен, как обычно, путем маргинализации — суммирования по значениям переменных S и W, причем суммирование по последней фактически тривиально из-за нормированности матрицы $P(W \mid R)$.

Значения апостериорных вероятностей и для дождя, и для поливальной машины выше соответствующих величин 0.3 и 0.2 для априорных вероятностей, как и следовало ожидать. Если теперь Холмс обнаружит, что трава у дома Ватсона влажная, то вероятности снова изменятся:

$$P(R = t \mid C = t, W = t) = \frac{P(R = t, C = t, W = t)}{P(C = t, W = t)} =$$
$$= \frac{0.8 \cdot 0.3 \cdot (0.18 + 0.64)}{0.8 \cdot (0.054 + 0.192) + 0.1 \cdot (0.098 + 0.056)} = \frac{0.1968}{0.2122} = 0.9274\,,$$
$$P(S = t \mid C = t, W = t) = 0.2498\,.$$

Ясно, что вероятность дождя возросла вследствие дополнительной информации о влаге на траве у дома Ватсона. Так как высокая вероятность дождя

объясняет влажность травы у дома самого Холмса, то объяснений при помощи другой причины (т. е. включенной поливальной установки) больше не требуется и вероятность ее понижается почти до исходного значения 0.2. Это и есть пример «попутного объяснения», о котором говорилось в предыдущих параграфах.

В общем случае при росте числа переменных в сети, задача точного нахождения вероятностей в сети является крайне вычислительно сложной [8]. Это вызвано комбинаторным сочетанием значений переменных в суммах при вычислении маргиналов от совместного распределения вероятностей, а также потенциальным наличием нескольких путей, связывающих пару переменных на графе. На практике часто используются приближенные методы для оценок комбинаторных сумм, например вариационные методы и многочисленные вариации методов Монте-Карло. Одним из таких приближенных подходов является метод выборок из латинских гиперкубов [9].

Выборочное оценивание вероятностей на латинских гиперкубах

Выборки из латинских гиперкубов начали широко использоваться после удачных решений в области планирования эксперимента, где их применение позволяет уменьшить взаимную зависимость факторов без увеличения числа экспериментов [19].

Представим вначале, что нам требуется разместить 8 фигур на 64 клетках шахматной доски так, чтобы в одномерных проекциях на вертикаль и горизонталь фигуры были распределены максимально равномерно. Решение этой классической задачи дается 8 ладьями, расположенными так, что, в соответствии с шахматными правилами, ни одна из них не бьет другую. Любое из 8! решений годится для приведенных условий. Однако, если мы добавим требование, чтобы двумерное распределение на доске было как можно ближе к равномерному, не все решения окажутся пригодными. В частности, решение, в котором все ладьи выстроились на одной диагонали, существенно недооценивает области двух противоположных углов.

Более удовлетворительное решение можно получить, многократно меняя местами строки для пар наудачу выбираемых ладей. При этом отбираются конфигурации, удовлетворяющие некоторым специальным услови-

[8] Эта задача относится к классу NP-полных. (*Cooper G. F.* The computational complexity of probabilistic inference using Bayesian belief networks // *Artificial Intelligence*. 1990. – **42**, (2–3). – pp. 393–405).

[9] LHS — Latin Hypercube Sampling.

ям — например, распределения, в которых минимальное расстояние между всеми парами ладей максимально, либо конфигурации с ортогональными [10] столбцами. Такой поиск соответствует перестановкам во множестве чисел $1, 2, \ldots, 8$, которые нумеруют строки для упорядоченных по столбцам фигур. Все эти размещения называются *латинскими квадратами*.

Латинские гиперкубы представляют собой прямое обобщение этого примера на случай N точек в пространстве D измерений. Все конфигурации получаются из базовой матрицы размещений размерности $(N \times D)$, в которой в каждом столбце последовательно расположены числа $1, 2, \ldots, N$.

Таким образом, метод латинских гиперкубов позволяет путем целочисленного комбинаторного перебора получить множество из N многомерных векторов, распределение которых, по построению, может обладать полезными дополнительными свойствами, в сравнении с полностью случайными векторами.

В применении к оцениванию условных вероятностей таким важным свойством может являться использование каждого из допустимых значений переменных-предков хотя бы один раз в выборке.

Вновь обратимся к численному примеру [4]. Пусть рассматривается байесова сеть из трех узлов с одним конвергентным соединением.

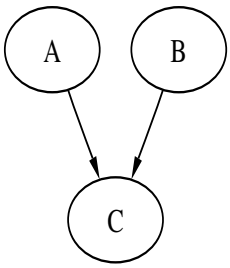

Рис. 3. Архитектура простейшей сети

В отличие от ранее рассмотренных примеров, переменные в сети могут принимать более чем два значения. Для проведения анализа вероятностей построим матрицу латинского гиперкуба (100×3), соответствующую 100 испытаниям. Эта матрица может выглядеть, например, так:

[10] Имеется в виду ортогональность столбцов при переходе к нумерации $-N/2 \ldots + N/2$.

ТАБЛИЦА 2. Архитектура и таблицы вероятностей в простейшей байесовой сети

A	$P(A)$
a_1	0.20
a_2	0.35
a_3	0.45

B	$P(B)$
a_1	0.4
a_2	0.6

	a_1		a_2		a_3	
$P(C \mid A, B)$	b_1	b_2	b_1	b_2	b_1	b_2
c_1	0.01	0.04	0.28	0.06	0.18	0.82
c_2	0.68	0.93	0.52	0.12	0.50	0.10
c_3	0.31	0.03	0.20	0.82	0.32	0.08

ТАБЛИЦА 3. Фрагмент латинского гиперкуба (100×3)

N	A	B	C
1	25	13	74
2	14	91	7
...	...	...	...
39	47	32	56
...	...	...	...
100	69	4	84

Числа от 1 до 100 соответствуют разбиению естественного вероятностного интервала $[0 \dots 1]$ на 100 равных частей. Пусть в i-м испытании для переменной A элемент матрицы гиперкуба равен LHS_iA. Выборка значений переменных далее проводится по следующим формулам:

$$\begin{cases} a_1\,, & \text{если} \quad LHS_iA \leqslant \big[P(a_1) \times 100)\big] \\ a_2\,, & \text{если} \quad \big[P(a_1) \times 100\big] \leqslant LHS_iA \leqslant \big[(P(a_1) + P(a_2)) \times 100\big] \\ a_1\,, & \text{в противном случае} \end{cases}$$

Аналогично, элемент LHS_iB служит для генерации значения переменной B. Пусть для генерации используется 39-я строка гиперкуба в табл. 3., в нашем случае $\{47, 32, 56\}$. Тогда переменная A принимает значение a_2, а B, соответственно, b_1.

Для получения C обратимся к матрице условных вероятностей $P(C \mid A, B)$. Комбинации (a_2, b_1) отвечает 3-й столбец, розыгрыш в котором по $LHS_iC = 56$ дает для C значение c_2. По этой же схеме разыгрываются значения всех переменных для всех 100 случаев. В итоге для переменной C накапливается выборочная гистограмма — оценка фактического распределения вероятности различных значений при заданных предположениях об остальных переменных.

В случае большего числа переменных и более сложной топологии сети, вычисления проводятся по цепочке, от предков к потомкам. К розыгрышу очередной переменной можно приступать, когда значения всех ее предков в данном примере выборки уже установлены [11].

Сложность вычислений пропорциональна числу переменных и размеру выборки, т. е. определяется размером гиперкуба, а не комбинаторным числом сочетаний значений переменных в сети.

Отметим, что метод латинского гиперкуба применим и для моделирования непрерывных распределений. В этом случае для розыгрыша безусловных вероятностей используется известный метод обратных функций распределения [24], а для интерпретации апостериорных вероятностей необходимо применить один из методов сглаживания плотности, заданной на дискретном множестве точек.

Замечание о субъективных вероятностях и ожиданиях

Исчисление вероятностей формально не требует, чтобы использованные вероятности базировались на теоретических выводах или представляли собой пределы эмпирических частот. Числовые значения в байесовых сетях могут быть также и субъективными, личностными, оценками ожиданий экспертов [12] по поводу возможности осуществления событий. У разных лиц степень ожидания (надежды или боязни — по *Лапласу*) события может

[11] Эта схема вычислений пригодна для любых ориентированных графов без направленных петель.

[12] Последовательное искоренение субъективных «вероятностей» имеет уходящие глубоко корнями в прошлое традиции в русской литературе о теории вероятностей. Изданная недавно энциклопедия «Вероятность и математическая статистика» (БРЭ, 1999) цитирует статью выдающегося математика *А. Н. Колмогорова* из энциклопедии 1951 года относительно субъективного понимания вероятности: «Оно приводит к абсурдному утверждению, что из чистого незнания, анализируя лишь одни субъективные состояния нашей большей или меньшей уверенности, мы сможем сделать какие-либо определенные заключения относительно внешнего мира» (ВИМС, с. 97).

быть разной, это зависит от индивидуального объема априорной информации и индивидуального опыта. Можно ли придать этим значениям ожиданий смысл вероятностей — полемический вопрос, выходящий далеко за рамки данной лекции.

Предложен оригинальный способ количественной оценки субъективных ожиданий. Эксперту, чьи ожидания измеряются, предлагается сделать выбор в игре с четко статистически определенной вероятностью альтернативы — поставить некоторую сумму на ожидаемое событие, либо сделать такую же ставку на событие с теоретически известной вероятностью (например, извлечение шара определенного цвета из урны с известным содержанием шаров двух цветов). Смена выбора происходит при выравнивании степени ожидания эксперта и теоретической вероятности. Теперь об ожидании эксперта можно (с небольшой натяжкой) говорить как о вероятности, коль скоро оно численно равно теоретической вероятности некоторого другого статистического события.

Использование субъективных ожиданий в байесовых сетях является единственной альтернативой на практике, если необходим учет мнения экспертов (например, врачей или социологов) о возможности наступления события, к которому неприменимо понятие повторяемости, а также невозможно его описание в терминах совокупности элементарных событий.

Более подробно вопрос о предмете теории вероятностей с точки зрения отечественной школы изложен, например, в [7], байесов подход к понятию вероятности представлен в [14].

Синтез и обучение байесовых сетей

В предыдущем разделе был подробно рассмотрен вопрос статистического вывода суждения о вероятностях различных значений переменных, которые в ряде случаев можно наблюдать на практике. Тем самым, решена задача вероятностного прогнозирования с использованием имеющихся представлений об априорных и условных вероятностях в байесовой сети.

При прогнозировании использовался байесов подход — априорные вероятности для распределения (части) переменных в сети в свете наблюдаемой информации приводят к оценкам апостериорных распределений, которые и служат для прогноза.

В этом разделе дается краткий обзор подходов к построению самих байесовых сетей: синтезу архитектуры и оценке параметров вероятност-

ных матриц для выбранной архитектуры. При обучении мы также будем опираться на последовательную байесову логику, в которой значения параметров также считаются случайными, а вместо поиска универсального единственного значения для каждого параметра оценивается его распределение. При этом, априорные предположения об этом распределении уточняются при использовании экспериментальных данных.

Проблема автоматического выбора архитектуры связей сети заметно сложнее задачи оценивания параметров. Дополнительные сложности возникают при использовании для обучения экспериментальных данных, содержащих пропуски. В самом сложном варианте должны вероятностно моделироваться и изменения архитектуры, и параметры условных вероятностей, и значения пропущенных элементов в таблицах наблюдений [13].

Синтез сети на основе априорной информации

Как уже отмечалось, вероятности значений переменных могут быть как физическими (основанными на данных), так и байесовыми (субъективными, основанными на индивидуальном опыте). В минимальном варианте полезная байесова сеть может быть построена с использованием только априорной информации (экспертных ожиданий).

Для синтеза сети необходимо выполнить следующие действия:

- сформулировать проблему в терминах вероятностей значений целевых переменных;
- выбрать понятийное пространство задачи, определить переменные, имеющие отношение к целевым переменным, описать возможные значения этих переменных;
- выбрать на основе опыта и имеющейся информации априорные вероятности значений переменных;
- описать отношения «причина-следствие» (как косвенные, так и прямые) в виде ориентированных ребер графа, разместив в узлах переменные задачи;
- для каждого узла графа, имеющего входные ребра указать оценки вероятностей различных значений переменной этого узла в зависимости от комбинации значений переменных-предков на графе.

[13] Эта задача, очевидно, является некорректно поставленной, как и всякая задача обучения. Описанный уровень сложности отражает необходимость глубокой регуляризации для согласования распределений такого большого числа ненаблюдаемых переменных.

Эта процедура аналогична действиям инженера по знаниям при построении экспертной системы в некоторой предметной области. Отношения зависимости, априорные и условные вероятности соответствуют фактам и правилам в базе знаний ЭС.

Построенная априорная байесова сеть формально готова к использованию. Вероятностные вычисления в ней проводятся с использованием уже описанной процедуры маргинализации полной вероятности.

Дальнейшее улучшение качества прогнозирования может быть достигнуто путем обучения байесовой сети на имеющихся экспериментальных данных. Обучение традиционно разделяется на две составляющие — выбор эффективной топологии сети, включая, возможно, добавление новых узлов, соответствующих скрытым переменным, и настройка параметров условных распределений для значений переменных в узлах.

Обучение байесовых сетей на экспериментальных данных

Если структура связей в сети зафиксирована, то обучение состоит в выборе свободных параметров распределений условных вероятностей $P[x_j \mid pa(x_j)]$. Для случая дискретных значений переменных x_j, в основном рассматриваемого в лекции, неизвестными параметрами являются значения матричных элементов распределений (см., например, табл. 1).

Максимально правдоподобными оценками этих матричных элементов могут служить экспериментальные частоты реализации соответствующих наборов значений переменных. Вспомним, однако, ситуацию с любительницей чая со сливками и профессиональным музыкантом, описанную во вводной части лекции. Статистические частоты, абсолютизируемые в методе максимального правдоподобия, не учитывают различия в экспертном опыте для разных экспериментальных ситуаций.

Для учета априорных знаний эксперта в вероятностном моделировании обратимся к байесову подходу к обучению. Рассмотрим сначала оба подхода — классический и байесов на примере простой задачи [14] обучения однопараметрической модели.

Пусть в нашем распоряжении имеется монета, свойства которой нам достоверно не известны. Проведем N экспериментов по подбрасыванию

[14] Описание этой задачи, восходящей еще к классикам, можно в том или ином виде найти во многих работах по теории вероятности. Но задача эта столь показательна, что комментарий к ней кажется просто необходимым при иллюстрации особенностей байесова подхода.

монеты и обозначим число выпавших «орлов» как h, а «решек», соответственно, $t = N - h$. Совокупность всех N наблюдений обозначим символом D. Зададимся теперь целью оценить вероятность различных исходов в следующем, $N + 1$ испытании.

В классическом подходе моделируемая система описывается одним фиксированным параметром — вероятностью выпадения «орла» θ, единственное («правильное») значение которого нужно оценить из эксперимента. При использовании метода максимального правдоподобия эта наилучшая оценка, очевидно, равна относительной частоте появления орлов в N экспериментах.

В байесовом подходе значение параметра θ само является *случайной* величиной, распределение которой используется при прогнозировании исхода следующего бросания монеты. Априорная (учитывающая наш опыт ξ до проведения испытаний) плотность распределения θ есть $p(\theta \mid \xi)$. После проведения серии экспериментов наши представления об этом распределении изменятся, в соответствии с теоремой Байеса:

$$p(\theta \mid D, \xi) = \frac{p(D \mid \theta, \xi) \cdot p(\theta \mid \xi)}{p(D \mid \xi) \triangleq \int p(D \mid \theta, \xi) \cdot p(\theta \mid \xi) d\theta} .$$

Функция правдоподобия, разумеется, одна и та же и в классическом, и в байесовом подходе, и равна биномиальному распределению:

$$p(D \mid \theta, \xi) \sim \theta^h \cdot (1 - \theta)^t$$

Прогноз вероятности будущего эксперимента дается формулой суммирования:

$$p(x_{N=1} = H \mid D, \xi) = \int p(x_{N+1} = H \mid \theta, \xi) \cdot p(\theta \mid D, \xi) d\theta =$$

$$= \int \theta \cdot p(\theta \mid d, \xi) d\theta = \langle \theta \rangle_{p(\theta|D,\xi)} .$$

Для получения интерпретируемого результата в замкнутом виде ограничим выбор априорных распределений классом β-*распределений* [14]:

$$p(\theta \mid \xi) = \beta(\theta \mid \alpha_h, \alpha_t) \triangleq \frac{\Gamma(\alpha_h + \alpha_t)}{\Gamma(\alpha_h) \cdot \Gamma(\alpha_t)} \theta^{\alpha_h - 1} \cdot (1 - \theta)^{\alpha_t - 1} .$$

Произведение двух биномиальных распределений вновь дает биномиальный закон, и это проясняет суть использования β-распределения в качестве априорного [15].

$$p(x_{N+1} = H \mid D, \xi) = \frac{\alpha_h + h}{\alpha_h + \alpha + t + h + t}\,.$$

Идея состоит в формализации опыта с бросанием монет путем добавления «искусственных» (полученных в гипотетических предыдущих экспериментах) отсчетов α_h «орлов» и α_t «решек» в экспериментальную серию. Чем больше мы добавим в экспериментальную выборку этих априорных наблюдений, тем меньше наша оценка вероятности $N+1$ испытания будет чувствовать возможные аномальные «выбросы» во множестве D. Поэтому байесово обучение иногда называют *обучением с априорной регуляризацией*.

Задача с одним параметром напрямую обобщается на обучение многопараметрической байесовой сети. Вместо бросания монеты генерируется случайный вектор, составленный из всех параметров сети, при этом исходы значений разыгрываются в соответствии с распределением Дирихле (обобщающем биномиальное распределение на случай более чем двух исходов). Теперь исходом испытания будет реализация значений вектора переменных в байесовой сети.

Если $D = \{D_1, \ldots, D_k, \ldots, D_S\}$ — множество обучающих примеров (каждый элемент D_k является вектором значений всех переменных сети в k-м примере), не содержащее пропущенных значений, то классический вариант обучения состоит в максимизации правдоподобия данных, как функции матричных элементов:

$$L = \frac{1}{N \cdot S} \sum_{j=1}^{N} \sum_{k=1}^{S} \log\bigl[P(x_j \mid pa(x_j), D_k)\bigr]\,.$$

Легко видеть, что обучение в этом подходе состоит в подсчете статистики реализаций векторов ситуаций для каждого матричного элемента в таблицах условных вероятностей. Максимально правдоподобными (наименее противоречащими экспериментальным данным) будут значения вероятностей, равные нормированным экспериментальным частотам.

[15] Априорные распределения, которые в итоге приводят к апостериорным распределениям из того же класса, называют *сопряженными* (conjugate priors).

Классическая схема максимального правдоподобия весьма проста, но ее основным недостатком в многомерном случае являются нулевые оценки вероятностей сочетаний значений переменных, не встретившихся в выборке. Между тем, из-за комбинаторного роста числа таких сочетаний в матрицах условных вероятностей будет все больше нулевых значений, а остальные значения будут сильно зашумлены из-за малого числа отсчетов. В байесовом подходе нулевые значения заменятся априорными вероятностями, которые позволят улучшить оценки прогнозируемых значений, но зато будут нивелированы экспериментальные свидетельства.

Таким образом, оба подхода имеют свои недостатки и преимущества, что и дает пищу нескончаемым дискуссиям сторонников классических и байесовых вероятностей. Практика требует примирения, но в настоящее время равновесие, по-видимому, смещено в сторону классического подхода [16].

Сложная задача автоматического синтеза топологии связей сети также обычно рассматривается на основе и классического, и байесового подходов. После внесения пробного изменения в топологию сети, два структурных варианта сравниваются путем оценки значения знаменателя формулы Байеса (evidence) или правдоподобия всей совокупности обучающих примеров. Далее архитектуры с лучшими значениями ценности отбираются в соответствии с принятой схемой рандомизированного или детерминированного поиска.

Обсудим здесь упрощенный вариант задачи синтеза топологии вероятностной сети для случая бинарного дерева, аппроксимирующего распределение условной вероятности для нескольких независимых и одной зависимой переменной. Обучение таких регрессионных деревьев может использоваться для сравнительного анализа значимости независимых переменных при их отборе для байесовой сетевой модели.

Вероятностные деревья

В этом разделе предлагается (относительно новый) подход последовательного упрощения задачи аппроксимации условной вероятности, основанный на вероятностных деревьях [10, 12]. Рассмотрим матрицу эмпирических

[16] Это — априорная оценка автора, который относит себя, скорее, к сторонникам байесового подхода. Классический подход популярен среди физиков [1], байесовы вероятности — среди специалистов по обучению машин и анализу данных (`www.auai.org`).

данных D, строки которой соответствуют реализациям случайных переменных $\{\vec{X}; Y\}$. Эти данные содержат информацию об условной плотности вероятности $p\,(y \mid \vec{x})$. Нашей целью будет построение функциональной модели этой плотности.

Для простоты ограничимся случаем скалярной зависимой переменной [17]. Если последняя принимает непрерывный ряд значений на некотором отрезке, то задача аппроксимации плотности соответствует задаче (нелинейной) регрессии. При этом независимые переменные могут быть как непрерывными, так и дискретными (упорядоченными или неупорядоченными).

При построении модели условной плотности мы будем опираться на байесов подход вывода заключений из данных (*reasoning*). А именно, имеющуюся информацию о значениях входных переменных обученная машина переводит в апостериорную информацию о характере распределения выходной (зависимой) переменной.

Процесс вывода суждения может носить итерационный характер: машина последовательно использует информацию, содержащуюся во входах для построения сужающихся приближений к апостериорной плотности [18].

Вопрос заключается в эффективном способе синтеза (обучения) такой вероятностной машины на основе лишь имеющихся эмпирических данных.

Метод построения связей и выбора правил в узлах дерева

Остановимся на простой информационной трактовке обучения искомой вероятностной машины. Для этого на множестве значений зависимой переменной введем дискретизацию $y_{\min} < y_1 < \ldots < y_k < \ldots y_{\max}$ таким образом, что в каждый отрезок попадает одинаковое число наблюдений $n_k = p_k \cdot N_0$ из матрицы наблюдений D. Тогда *априорная* заселенность [19] всех интервалов одинакова, что соответствует максимальной энтропии:

$$S_0 = -N_0 \sum_k p_k \log p_k \,.$$

[17] Это равносильно обычному предположению о взаимной независимости зависимых переменных *при условии*, что значения независимых переменных заданы.

[18] В качестве априорного распределения у используется частотная гистограмма значений по совокупности данных D. Путем предобработки данных это распределение приводится к константе (т. н. неинформативный prior).

[19] Экспериментальные частоты в отсутствие другой информации являются наиболее вероятной оценкой *априорной* плотности вероятности.

Это (максимальное) значение энтропии отвечает полному отсутствию информации о возможном значении зависимой переменной. Здесь использовано предположение о независимости и одинаковом (стационарном) характере распределения всех отдельных наблюдений в матрице данных.

Построим теперь такой процесс вероятностного вывода, при котором дополнительная информация о входных (независимых) переменных приводит к уменьшению энтропии распределения выходной переменной.

Прямой путь состоит в генерации двоичного дерева, в каждом узле которого применяется простейшее решающее правило. Применение этого правила разделяет исходную совокупность данных на два множества. Идея заключается в том, чтобы суммарная энтропия распределений $S_1 = S_1^{'} + S_1^{''}$ в полученных множествах была меньше исходной. Другими словами, оптимальное решающее правило выбирается таким образом, чтобы его *информативность* была максимальной, и, соответственно, остаточная энтропия после его применения — минимальной [20].

Ограничимся в этой работе простым классом правил, в которых значение одной из зависимых переменных сравнивается с порогом. Правила такого типа широко применяются в классификаторах на основе бинарных деревьев [12].

Оптимальное (на данном уровне иерархии) правило выбирается в процессе последовательного решения M *одномерных* задач оптимизации (M — число независимых переменных X). Целевая функция — суммарная энтропия двух подмножеств, получаемых при применении правила, изменяемая переменная — значения порога для правила. Выбор останавливается на правиле, обеспечившем максимальное уменьшение энтропии.

Для дискретных переменных вместо решения задачи гладкой оптимизации выполняется перебор возможных значений классов с решающим правилом «свой класс—все остальные».

Иерархический процесс далее продолжается для подмножеств (дочерних ветвей и соответствующих им примеров) данного узла дерева. Процесс формально завершается по достижении нулевой энтропии для каждого узла самого нижнего уровня (значение зависимой переменной для всех примеров данного узла попадает в один интервал исходной дискретизации).

В итоге, каждому узлу полученного дерева приписывается:

[20] Заметим, что традиционно используются другие критерии выбора правил при построении деревьев, в частности, минимизируется дисперсия зависимой переменной, либо какие-то другие функционалы.

1. Эмпирическая оценка плотности условного распределения дискретизованной зависимой переменной (при условии отнесения примера к данному узлу).
2. Оценка выборочной энтропии распределения в этом узле.
3. Решающее правило, позволяющее выбрать дочернюю ветвь с дальнейшим уменьшением энтропии условного распределения.

Свойства вероятностного дерева

Обученная по предлагаемой методике машина предлагает в ответ на информацию о векторе независимых переменных целую серию последовательно уточняющихся приближений к оценке апостериорной плотности условного распределения зависимой переменной.

Обобщающая способность. В таком максимальном варианте дерево является, очевидно, *переобученным*, так как в нем полностью запомнен весь шум, содержавшийся в данных. На практике, иерархия предсказаний плотности может быть остановлена на более ранних уровнях (например, в узле должно содержаться не менее определенного числа примеров обучающей выборки, либо энтропия распределения должна быть выше порогового значения).

Для оценок можно воспользоваться методикой кросс-валидации на основе бутстрэп-выборок [25]. Эти простые эмпирические способы регуляризации могут быть заменены более последовательными методами минимизации длины описания модели и данных [2].

Сходимость метода. Нетрудно убедиться, что для любого конечного набора данных размера N, после отщепления одного примера суммарная энтропия полученных двух множеств (отщепленный пример и совокупность остальных примеров) строго меньше энтропии исходного набора. Действительно, энтропия отдельного примера равна 0, а энтропия оставшихся примеров лимитируется множителем $(N-1)/N$.

Тем самым, метод всегда сходится, по крайней мере за $(N-1)$ шагов. Практическая скорость убывания энтропии в типичных вычислительных экспериментах приведена на рис. 4.

При этом вклад различных входных переменных в снижение энтропии весьма неоднороден.

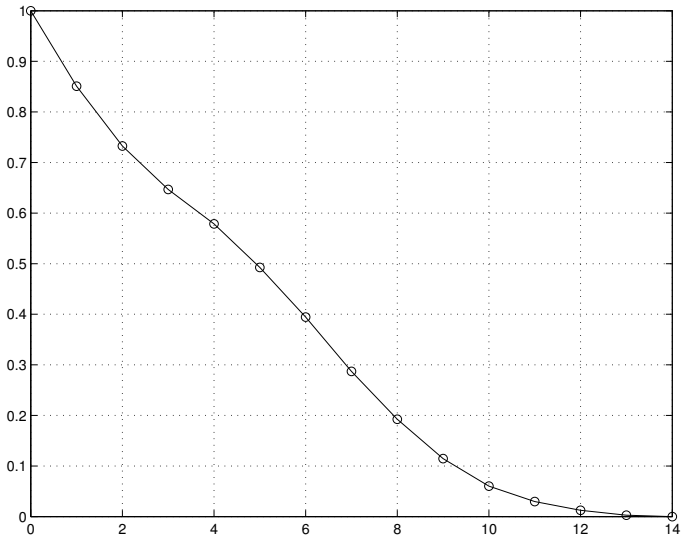

Рис. 4. Зависимость энтропии (в долях к исходному значению) от уровня иерархии дерева

Значимость независимых переменных (факторов). Важным «побочным» эффектом методики является возможность оценки относительной значимости факторов-входов. Все факторы (независимые переменные) можно упорядочить по степени уменьшения суммарной энтропии зависимой переменной, возникшей при использовании фактора в решающих правилах. Не использованные ни в одном из правил переменные в этом подходе признаются *незначимыми*.

Характерное распределение значимости факторов приведено на рис. 5.

Заметим, что предлагаемый метод приводит к весьма разреженному представлению модели данных, что может оказаться полезным при увеличении числа независимых переменных.

Вычислительная сложность метода. Применение метода требует решения множества задач *одномерной* оптимизации на отрезке. При этом затраты на вычисление оптимизируемой функции падают линейно по мере продвижения от корня дерева к листьям.

Задачи безусловной оптимизации функций на отрезке весьма подробно исследованы, и соответствующие методики успешно реализованы в пакетах прикладных программ [8].

Вычислительные затраты на одномерную оптимизацию несравненно меньше затрат на многомерный поиск, который требуется, например, при построении статистических моделей EM-алгоритмом, или обучении многослойных нейронных сетей [2].

Поиск решающих правил в ветвях дерева может быть непосредственно распараллелен, что повышает его эффективность при использовании современных ЭВМ.

Аналоги метода. Близким аналогом вероятностного дерева является *нейросетевая архитектура со встречным распространением* (counter-propagation, *R. Hecht-Nielsen*, [13]).

На кластеризующем [16] слое нейросеть со встречным распространением, являющаяся, по существу, одноуровневым деревом, происходит сквозной просмотр всех ветвей-кластеров. При этом, в отличие от предлагаемого подхода, для выбора ветви дерева используется метрика в пространстве входов вместо решающего правила. Таким образом, кроме повышенных вычислительных затрат, на практике могут возникать трудности с выбором подходящей метрики, что не всегда возможно в случае дискретных и непрерывных входов. Заметим также, что кусочно-постоянная аппроксимация в форме ячеек Вороного в нейросети со встречным распространением качественно отличается от глобального представления функции в виде иерархических сегментов постоянства функции при использовании дерева.

О применениях вероятностных деревьев

Оценка стоимости недвижимости в районе Бостона. В этой, ставшей классической [11,23], задаче предлагается построить прогностическую систему оценки стоимости недвижимости в районе Большого Бостона. База данных [11] содержит 506 наблюдений для 13 зависимых и одной независимой переменной (собственно цены).

Все переменные в базе данных представлены действительными числами, за исключением одной двоичной переменной (признак — находится ли строение в районе реки Чарльз Ривер)

Построение модели дерева заняло чуть больше минуты на персональной ЭВМ. Полное исчерпание энтропии в данных достигнуто при 439 узлах (число выведенных решающих правил чуть больше 200).

Относительная значимость факторов представлена на рис. 5.

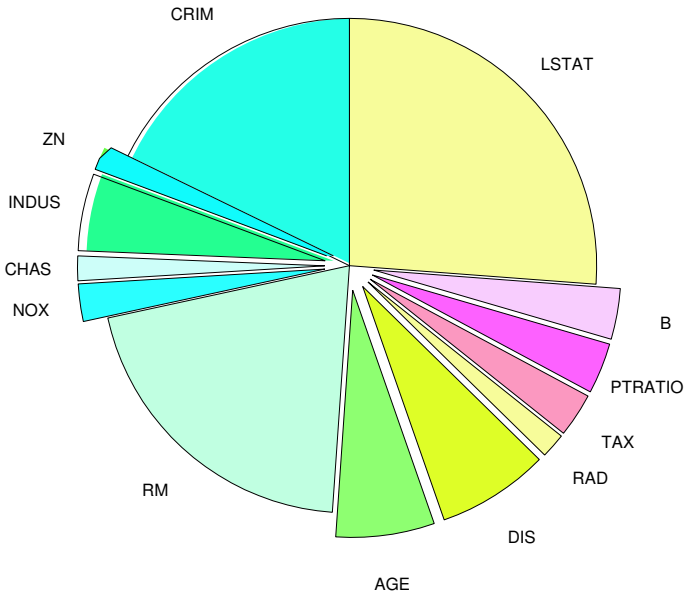

Рис. 5. Относительная значимость факторов в задаче Boston Housing

Почти $3/4$ всей неопределенности в прогнозе цены устраняется на основе информации лишь о трех «прозаических» факторах: ZN — процент земельной площади, выделенной для больших участков (свыше 25000 кв. футов), CRM — уровень преступности, RM — число комнат в доме.

Характер последовательного сужения плотности распределения при перемещении вниз по уровням иерархии дерева показан на рис. 6.

Прогноз свойств смеси химических компонент. Одним из успешных практических применений модели вероятностных деревьев явилось решение задачи о прогнозе свойств смеси химических красителей, применяемых в бытовой химии.

Особенность задачи состоит в том, что около полутора десятков химических компонентов вступают в сложные химические реакции, выход которых нелинейно зависит от концентрации веществ, при этом характер процессов является вероятностным. Последнее обстоятельство приводит к наличию шума и (статистических) противоречий в данных.

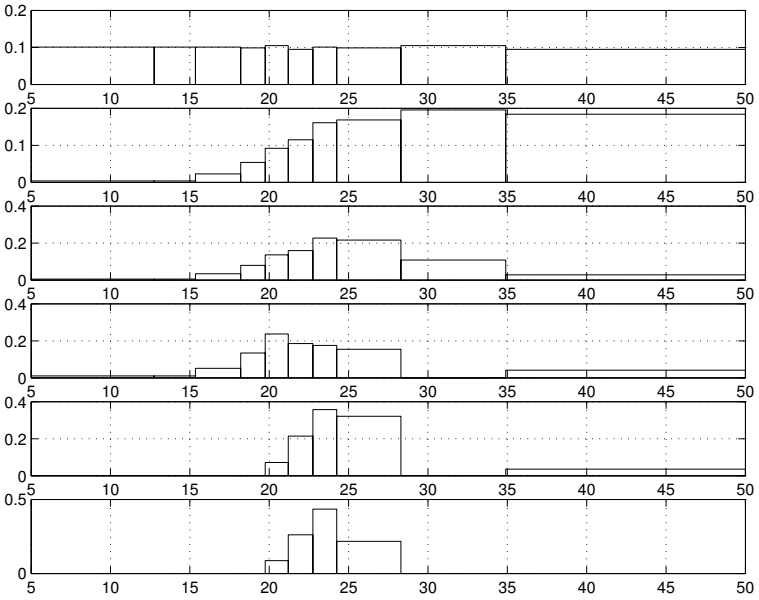

Рис. 6. Упорядочение плотности распределения цены при перемещении по иерархии дерева

Применение вероятностных деревьев позволило построить компактные прогностические модели свойств смеси.

Замечания о компьютерной реализации вероятностных деревьев. Проведенные вычислительные эксперименты показывают высокую эффективность и простоту применения обучающихся машин на основе вероятностных деревьев.

Основные применения полученных моделей состоят в анализе значимости входных факторов и возможности вероятностного прогноза значений выходных переменных при известных входах.

Приведение распределения выходов к константе не является обязательным, и выполнено в примерах здесь лишь для чистоты иллюстрации метода.

В предлагаемом подходе одинаково легко анализируются континуальные и дискретные входы. Малорелевантные входы автоматически игнорируются, что не приводит к усложнению методики.

Примеры применений байесовых сетей

Естественной областью использования байесовых сетей являются экспертные системы, которые нуждаются в средствах оперирования с вероятностями. Назовем лишь несколько областей с успешными примерами практических применений байесовых сетей. Некоторые коммерческие разработки и программное обеспечение обсуждаются в Приложении.

Медицина

Система PathFinder (Heckerman, 1990) разработана для диагностики заболеваний лимфатических узлов. PathFinder включает 60 различных вариантов диагноза и 130 переменных, значения которых могут наблюдаться при изучении клинических случаев. Система смогла приблизиться к уровню экспертов, и ее версия PathFinder-4 получила коммерческое распространение.

Множество других разработок (Child, MUNIN, Painulim, SWAN и др.) успешно применяются в различных медицинских приложениях [15].

Космические и военные применения

Система поддержки принятия решений Vista (*Eric Horvitz*) применяется в Центре управления полетами NASA (NASA Mission Control Center) в Хьюстоне. Система анализирует телеметрические данные и в реальном времени идентифицирует, какую информацию нужно выделить на диагностических дисплеях.

В исследовательской лаборатории МО Австралии системы, основанные на байесовых сетях, рассматриваются, как перспективные в тактических задачах исследования операций. В работе [9] описано применение пакета Netica (см. Приложение) к решению учебной задачи охраны территориальной зоны с моря “Operation Dardanelles”. Модель включает в себя различные тактические сценарии поведения сторон, данные о передвижении судов, данные разведнаблюдений и другие переменные. Последовательное

поступление информации о действиях противников позволяет синхронно прогнозировать вероятности различных действий в течение конфликта.

Компьютеры и системное программное обеспечение

В фирме Microsoft методики байесовых сетей применены для управления интерфейсными агентами-помощниками в системе Office (знакомая многим пользователям «скрепка»), в диагностике проблем работы принтеров и других справочных и wizard-подсистемах.

Обработка изображений и видео

Важные современные направления применений байесовых сетей связаны с восстановлением трехмерных сцен из двумерной динамической информации, а также синтеза статических изображений высокой четкости из видеосигнала (*Jordan*, 2002).

Финансы и экономика

В серии работ школы бизнеса Университета штата Канзас [22] описаны байесовы методики оценки риска и прогноза доходности портфелей финансовых инструментов. Основными достоинствами байесовых сетей в финансовых задачах является возможность совместного учета количественных и качественных рыночных показателей, динамическое поступление новой информации, а также явные зависимости между существенными факторами, влияющими на финансовые показатели.

Результаты моделирования представляются в форме гистограмм распределений вероятностей, что позволяет провести детальный анализ соотношений «риск–доходность». Весьма эффективными являются также широкие возможности по игровому моделированию.

Обсуждение

Байесовы вероятностные методы обучения машин являются существенным шагом вперед, в сравнении с популярными моделями «черных ящиков». Они дают понятное объяснение своих выводов, допускают логическую интерпретацию и модификацию структуры отношений между переменными

задачи, а также позволяют в явной форме учесть априорный опыт экспертов в соответствующей предметной области.

Благодаря удачному представлению в виде графов, байесовы сети весьма удобны в пользовательских приложениях.

Байесовы сети базируются на фундаментальных положениях и результатах теории вероятностей, разрабатываемых в течение нескольких сотен лет, что и лежит в основе их успеха в практической плоскости. Редукция совместного распределения вероятностей большого числа событий в виде произведения условных вероятностей, зависящих от малого числа переменных, позволяет избежать «комбинаторных взрывов» при моделировании.

Байесова методология, в действительности, шире, чем семейство способов оперирования с условными вероятностями в ориентированных графах. Она включает в себя также модели с симметричными связями (случайные поля и решетки), модели динамических процессов (марковские цепи), а также широкий класс моделей со скрытыми переменными, позволяющих решать вероятностные задачи классификации, распознавания образов и прогнозирования.

Благодарности

Автор благодарит своих коллег по работе *С. А. Шумского*, *Н. Н. Федорову*, *А. В. Квичанского* за плодотворные обсуждения. Отдельная благодарность *Ю. В. Тюменцеву* за высококачественную редакторскую работу, важные замечания и его энтузиазм в отношении организации лекций, без которого это издание было бы невозможным. Спасибо родному МИФИ — *Alma Mater* — за организацию Школы по нейроинформатике и гостеприимство.

Литература

1. *D'Agostini G.* Bayesian reasoning in high energy physics — principles and applications. – CERN Yellow Report 99-03, July 1999.
2. *Bishop C.M.* Neural networks for pattern recognition. – Oxford University Press, 1995.
3. *Вентцель Е. С.* Теория вероятностей. – М. Высшая Школа, 2001.
4. *Cheng J., Druzdel M. J.* Latin hypercube sampling in Bayesian networks // In: *Proc. of the Thirteenth International Florida Artificial Intelligence Research Symposium*

(FLAIRS-2000). – Orlando, Florida, AAAI Publishers, 2000. – pp. 287–292.
URL: http://www2.sis.pitt.edu/~jcheng/Latin.zip

5. *Coles S.* Bayesian inference. Lecture notes. – Department of Mathematics, University of Bristol, June 10, 1999.
URL: http://www.stats.bris.ac.uk/~masgc/teaching/bayes.ps
6. *Giarratano J., Riley G.* Expert systems: Principles and programming. – PWS Publishing, 1998.
7. *Гнеденко Б. В.* Курс теории вероятностей. – 7-е изд. – М.: Эдиториал УРСС, 2001.
8. *Дэннис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988.
9. *Das B.* Representing uncertainties using Bayesian networks. – DSTO Electronics and Surveillance Research Laboratory, Department of Defence. – Tech. Report DSTO-TR-0918. – Salisbury, Australia, 1999.
URL: http://dsto.defence.gov.au/corporate/reports/DSTO-TR-0918.pdf
10. *Fukuda T., Morimoto Y., Morishita S., Tokuyama T.* Constructing efficient decision trees by using optimized numeric association rules // The VLDB Journal, 1996.
URL: http://citeseer.nj.nec.com/fukuda96constructing.html
11. *Harrison D., Rubinfeld D. L.* Hedonic prices and the demand for clean air // J. Environ. Economics & Management. – 1978. – vol. 5. – pp. 81–102.
12. *Hastie T., Tibshirani R., Friedman J.* The Elements of statistical learning. – Data Mining, Inference, and Prediction. Springer, 2001.
13. *Hecht-Nielsen R.* Neurocomputing. – Addison-Wesley, 1990
14. *Heckerman D.*. A tutorial on learning with Bayesian Networks. – Microsoft Tech. Rep. – MSR–TR–95–6, 1995.
15. *Jensen F. V.* Bayesian networks basics. – Tech. Rep. Aalborg University, Denmark, 1996.
URL: http://www.cs.auc.dk/research/DSS/papers/jensen:96b.ps.gz
16. *Kohonen T.* Self-organizing Maps. – Springer, 1995.
17. *Minka T.* Independence diagrams. – Tech. Rep. MIT, 1998.
URL: http://www-white.media.mit.edu/~tpminka/papers/diagrams.html
18. *Blake C. L., Merz C. J.* UCI Repository of Machine Learning Databases, 1998.
URL: http://www.ics.uci.edu/~mlearn/MLRepository.html
19. *Montgomery D. C.* Design and analysis of experiments. – 5th. Ed. – Wiley, 2001.
20. *Neal R. M.* Probabilistic inference using Markov chain Monte Carlo methods. – Technical Report CRG-TR-93-1, 25 Sep 1993, Dept. of Computer Science, University of Toronto.

21. *Pearl J.* Probabilistic reasoning in intelligent systems: Networks of plausible inference. – Morgan Kaufmann, 1988.

22. *Shenoy C., Shenoy P.* Bayesian network models of portfolio risk and return / Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo (Eds.) Computational Finance, 85-104, MIT Press, 1999.

23. StatLib datasets archive. – CMU.
URL: http://lib.stat.cmu.edu/datasets/

24. *Соболь И.М.* Численные методы Монте-Карло. – М.: Наука, 1973.

25. *Терехов С. А.* Нейросетевые аппроксимации плотности в задачах информационного моделирования. Лекция для школы-семинара «Современные проблемы нейроинформатики», Москва, МИФИ, 25–27 января 2002 года.

26. *Терехов С. А.* Нейросетевые информационные модели сложных инженерных систем. Глава IV в кн.: *А. Н.Горбань, В. Л.Дунин-Барковский, А. Н. Кирдин, Е. М. Миркес, А. Ю. Новоходько, Д. А. Россиев, С. А. Терехов, М. Ю. Сенашова, В. Г. Царегородцев*. Нейроинформатика. – Новосибирск, Наука, 1998ю – с. 101–136.

27. *Терехов С. А.* Лекции по теории и приложениям искусственных нейронных сетей. – Снежинск, 1994–1998.
URL: http://alife.narod.ru/lectures/neural/Neu_index.htm

Приложение А. Обзор ресурсов Интернет по тематике байесовых сетей

Байесовы сети — интенсивно развивающаяся научная область, многие результаты которой уже успели найти коммерческое применение. В этом Приложении представлены лишь немногие популярные ресурсы Интернет, спектр которых, однако, отражает общую картину, возникшую в последние 10–15 лет.

AUAI — Ассоциация анализа неопределенности в искусственном интеллекте

URL: http://www.auai.org/

Ассоциация анализа неопределенности в искусственном интеллекте (Association for Uncertainty in Artificial Intelligence — AUAI) — некоммерческая организация, главной целью которой является проведение ежегодной *Конференции по неопределенности в искусственном интеллекте* (UAI).

Конференция UAI–2002 прошла в начале августа 2002 года в Университете Альберты (Эдмонтон, Канада). Конференции UAI проходят ежегодно, начиная с 1985 года, обычно в совместно с другими конференциями по смежным проблемам. Труды конференций издаются в виде книг, однако многие статьи доступны в сети.

NETICA

URL: http://www.norsys.com/index.html)

Norsys Software Corp. — частная компания, расположенная в Ванкувере (Канада). *Norsys* специализируется в разработке программного обеспечения для байесовых сетей. Программа Netica — основное достижение компании, разрабатывается с 1992 года и стала коммерчески доступной в 1995 году. В настоящее время Netica является одним из наиболее широко используемых инструментов для разработки байесовых сетей.

Версия программы с ограниченной функциональностью свободно доступна на сайте фирмы *Norsys*.

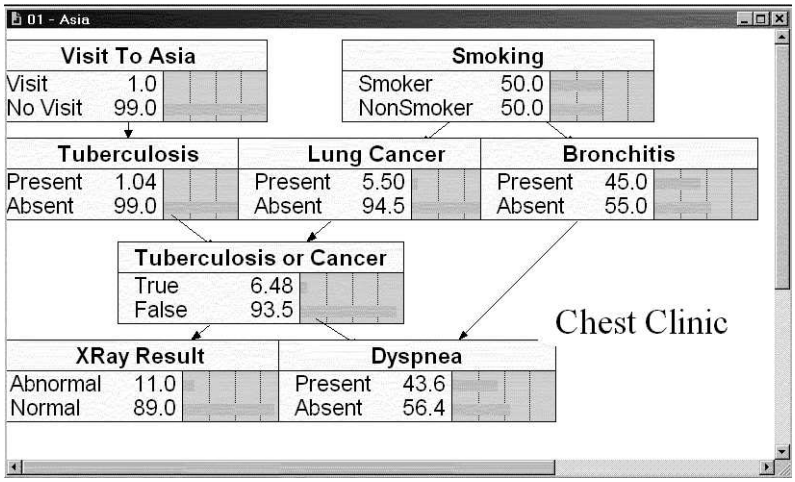

Рис. 7. Пример байесовой сети («Азия») в приложении Netica

Netica — мощная, удобная в работе программа для работы с графовыми вероятностными моделями. Она имеет интуитивный и приятный интерфейс пользователя для ввода топологии сети. Соотношения между переменными могут быть заданы,

как индивидуальные вероятности, в форме уравнений, или путем автоматического обучения из файлов данных (которые могут содержать пропуски).

Созданные сети могут быть использованы независимо, и как фрагменты более крупных моделей, формируя тем самым библиотеку модулей. При создании сетевых моделей доступен широкий спектр функций и инструментов.

Многие операции могут быть сделаны несколькими щелчками мыши, что делает систему Netica весьма удобной для поисковых исследований, и для обучения и для простого просмотра, и для обучения модели байесовой сети. Система Netica постоянно развивается и совершенствуется.

Knowledge Industries

URL: http://www.kic.com/

Knowledge Industries — ведущий поставщик программных инструментальных средств для разработки и внедрения комплексных диагностических систем. При проектировании сложных и дорогостоящих вариантов систем диагностик в компании используется байесовы сети собственной разработки.

Data Digest Corporation

URL: http://www.data-digest.com/home.html

Data Digest Corporation является одним из лидеров в применении методов байесовых сетей к анализу данных.

HUGIN Expert

URL: http://www.hugin.com/

Компания *Hugin Expert* была основана в 1989 году в Ольборге, (Дания). Их основной продукт Hugin начал создаваться во время работ по проекту *ESPRIT*, в котором системы, основанные на знаниях, использовались для проблемы диагностирования нервно-мышечных заболеваний. Затем началась коммерциализация результатов проекта и основного инструмента — программы Hugin. К настоящему моменту Hugin адаптирована во многих исследовательских центрах компании в 25 различных странах, она используется в ряде различных областей, связанных с анализом решений, поддержкой принятия решений, предсказанием, диагностикой, управлением рисками и оценками безопасности технологий.

BayesWare, Ltd

URL: http://www.bayesware.com/corporate/profile.html

Компания *BayesWare* основана в 1999 году. Она производит и поддерживает программное обеспечение, поставляет изготовленные на заказ решения, предоставляет программы обучения, и предлагает услуги консультирования корпоративным заказчикам и общественным учреждениям. Одна из успешных разработок компании, Bayesware Discoverer, основана на моделях байесовых сетей.

Персональные страницы специалистов по байесовым методам

URL: http://stat.rutgers.edu/∼madigan/bayes_people.html

На этой странице собраны адреса домашних страниц специалистов по байесовым сетям из лабораторий и фирм по всему миру.

Сергей Александрович ТЕРЕХОВ, кандидат физико-математических наук, заведующий лабораторией искусственных нейронных сетей Снежинского физико-технического института (СФТИ), заместитель генерального директора ООО «НейрОК». Область научных интересов — анализ данных при помощи искусственных нейронных сетей, генетические алгоритмы, марковские модели, байесовы сети, методы оптимизации, моделирование сложных систем. Автор 1 монографии и более 50 научных публикаций.