УДК 62-50

ПОСТРОЕНИЕ И МЕТОДЫ ОБУЧЕНИЯ БАЙЕСОВСКИХ СЕТЕЙ

Бидюк П.И., Терентьев А.Н., Гасанов А.С.

Введение. Байесовские сети (БС) представляют собой графические модели событий и процессов на основе объединения некоторых результатов теории вероятностей и теории графов. Они тесно связаны с диаграммами влияния, которые можно использовать для принятия решений. Несмотря на общепринятое название, эти сети не обязательно подразумевают тесную связь с байесовскими методами. Название связано, прежде всего, с байесовским правилом вероятностного вывода [1]. БС представляют собой удобный инструмент для описания достаточно сложных процессов и событий с неопределенностями структурного и статистического характера. Они оказались особенно полезными при разработке и анализе алгоритмов распознавания образов и ситуаций, анализе рисков различной природы, машинных алгоритмов обучения. Основной идеей построения графической модели есть понятие модульности, то есть разложение сложной системы на простые элементы в процессе анализа ее функционирования. При последующем объединении отдельных элементов в систему используются результаты теории вероятностей, которые обеспечивают состоятельность модели в целом, а также дают возможность объединять графические модели с базами данных. Такой подход к построению граф-теоретической модели обеспечивает исследователю возможность строить модели процессов, характеризующихся множеством сильно взаимодействующих переменных, а также создавать структуры данных, удобные для последующей разработки эффективных алгоритмов их обработки и принятия решений.

Формализм построения обобщенных графических моделей объединяет в себе много методов статистического моделирования, таких как факторный анализ, анализ распределений, модели смесей распределений, скрытые марковские модели, фильтры Калмана, модели Айзинга и некоторые другие. Все указанные методы и модели можно рассмотреть в рамках графических моделей байесовского типа как частные примеры общего формализма [2, 3, 4, 6]. Преимуществом такого подхода есть то, что методы исследования процессов и обработки данных, разработанные в одной области, могут быть успешно перенесены в другие.

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

Постановка задачи. Целью данной работы есть исследование общих принципов построения байесовских сетей, анализ методов их обучения и возможностей применения, а также иллюстрация применения теории на некоторых примерах.

1 Понятие байесовской сети

Графические модели представляют собой графы, узлы которых соответствуют случайным переменным. Если узлы (переменные) не соединены дугами, то их считают условно независимыми. Ненаправленные графические модели называют также марковскими случайными полями (МСП). Для МСП независимость можно сформулировать следующим образом: два множества (узлов) A и B являются условно независимыми при наличии в модели третьего множества C, если все пути между узлами множеств A и B разделены узлами множества C.

Байесовская сеть (БС) доверия определяется парой компонент – направленным ациклическим графом и множеством параметров, определяющими сеть. Первая компонента G является направленным ациклическим графом, соответствующим случайным переменным, которые используются для описания статики или динамики объекта.

БС используются для моделирования предметных областей, которые характеризуются структурной и статистической неопределенностью. Неопределенности возникают в случае, когда невозможно выбрать (оценить) модель с точной структурой или когда нет достаточной информации о состоянии процесса в момент принятия решения, когда механизмы, определяющие поведение системы, носят случайный характер или система функционирует в условиях комбинированного влияния упомянутых факторов. Вершины (узлы) БС соединяются направленными ребрами, при этом каждому узлу ставится в соответствие определенная вероятностная функция. Таким образом, БС – направленный ациклический граф (НАГ), который не содержит маршрута, начинающегося и заканчивающегося в одной и той же вершине.

Вершина БС представляет собой дискретную переменную с конечным числом состояний или непрерывную гауссову случайную величину. “Узел” и “вершина” употребляют как синонимы “переменной”. Ребра между вершинами представляют причинные связи между ними.

Вершина, не имеющая родительских узлов (то есть, ребер, направленных к ней) характеризуется таблицею безусловных вероятностей. В случае дискретной переменной эта таблица содержит распределение вероятностей между всеми возможными состояниями этой вершины. Если же вершина имеет родителей (то есть, одно или несколько ребер, направленных к ней), то она характеризуется таблицей условных вероятностей, содержащей значения условных вероятностей пребывания вершины в определенном состоянии в случае определенной конфигурации состояний ее родителей. Таким образом, число элементов таблицы условных вероятностей дискретной вершины БС равняется произведению чисел возможных состояний всех ее родительских вершин.

Простая тривиальная сеть, представленная на рис. 1, отражает причинно следственную связь между двумя элементами некоторой предметной области A и B. Существование причинной связи от A к B означает, что если A находится в некотором состоянии, это имеет влияние на состояние B.

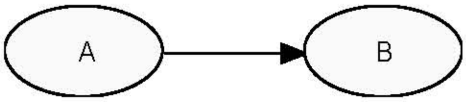


Рис. 1. Тривиальная байесовская сеть

Случайная дискретная переменная, соответствующая вершине *A,* может находиться в одном из двух состояний – a_1 или a_2. Вершина *B* имеет три возможных состоянитя: b_1, b_2, b_3. Таблицы условных вероятностей для этих вершин имеют следующий вид:

A	$P(a_i)$
A_1	0.5
A_2	0.5

B	$P(b_i \mid a_1)$	$P(b_i \mid a_2)$
b_1	1	0.6
b_2	0	0.2
b_3	0	0.2

Так как вершина *A* не имеет родителей, значения вероятностей ее состояний не зависят от других узлов. О распределении вероятностей вершины A в таком случае говорят, что заданы априорные вероятности ее состояний. Вероятности состояний вершины B, на оборот, зависят от состояний ее родительской вершины A:

- если A находится в состоянии a_1, то B находится в состоянии b_1;
- если A находится в состоянии a_2, то вероятность состояния b_1 для вершины B составляет 0,6, а состояний b_2 и b_3 – 0,2.

Таким образом, БС – это пара ***(G,P)***, де ***G*** = (*V*,*E*) – направленный ациклический граф на конечном множестве вершин *V*, соединенных между собой множеством направленных ребер *E*, а ***P*** – множество (условных) распределений вероятностей. БС имеет следующие свойства:

- каждая вершина из множества V может принимать одно значение из конечного множества взаимоисключающих значений (состояний);
- каждой вершине $A \in V$ с родительскими вершинами $B_1, B_2, \ldots B_n$ поставлена в соответствие таблица условных вероятностей $P(A \mid B_1, B_2, \ldots B_n)$.

Если вершина A не имеет родителей, то вместо условных вероятностей (автоматически) используются безусловные вероятности *P(A)*. Следует отметить, что требование отсутствия ориентированных циклов в графе является существенным, поскольку для направленных графов с циклами не предложены методы вычисления вероятностей.

Именно процесс вычисления вероятностей является основой для принятия решений на основе БС в условиях неопределенности. Раскрытие неопределенностей осуществляется в БС путем вычисления вероятностей состояний вершин на основе имеющейся информации о значениях (части) других вершин сети. Математической основой для этого служит байесовский поход к анализу неопределенностей и соответствующий ему аппарат классической теории вероятностей.

Основу байесовского подхода составляет понятие условной вероятности. Условная вероятность $P(A \mid B) = x$ означает, что при условии возникновения события B (а также всего юного, что не имеет отношения к B) вероятность события A равняется x. Совместная вероятность наступления событий A и B определяется по формуле полной вероятности:

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

Приведенное уравнение является фундаментальным правилом исчисления вероятностей и основой для теоремы Байеса:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}.$$

Теорема Байеса используется в случае, наличия информации о зависимых переменных (свидетельствах), а суть исследования состоит в том, чтобы определить вероятности исходных переменных (причин). Так, при наличии условной вероятности $P(B \mid A)$ возникновение некоторого события B при условии, что имеет место событие A, теорема Байеса позволяет решить обратную задачу – какой будет вероятность возникновения события A, если состоялось событие B.

Действительно, пусть $A_{1,} A_2, \ldots, A_n$ – полная группа несовместных взаимоисключающих событий (или альтернативных гипотез). Тогда *апостериорная* вероятность $P(A_j \mid B)$ каждого события $A_j, j = 1..n$ при условии, что состоялось событие *B*, выражается через априорную вероятность A_j:

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

Перед тем, как дать описание практических алгоритмов исчисления вероятностей в байесовских сетях, приведем определения некоторых понятий, характерных для теории БС.

- *Совместная вероятность* (совместное распределение вероятностей) всех вершин сети является наиболее полным статистическим описанием наблюдений. Совместное распределение представляется функцией многих переменных в соответствии с числом переменных конкретной задачи. В общем случае такое описание требует задания вероятностей всех допустимых конфигураций состояний всех переменных, что не представляется возможным для практической реализации. Байесовская сеть представляет собой компактное представление полной вероятности – концепцию сохранения информации, на основе которой может быть вычислена полная вероятность.

- *Цепное правило* является средством вычисления полной вероятности в БС. Так, если БС – байесовская сеть на множестве вершин $U = \{A_1, A_2, \ldots, A_n\}$, то совместное распределение вероятностей является произведением всех условных вероятностей, определенных в БС:

➢

$$P(U)=\prod_i P(A_i\,|\,par(A_i)),\qquad (4)$$

где $par(A_i)$ – множество (состояний) родительских вершин для A_i.

➢ *Условная независимость* вершин байесовской сети означает блокирование взаимного влияния между этими вершинами. Переменные (множества переменных) A и C являются независимыми при известном состоянии переменной B, если

$$P(A\,|\,B)=P(A\,|\,B,C).\qquad (5)$$

Это означает, что если состояние вершины B известно, то никакая информация о C не изменит вероятностей A.

➢ *d*-разделимость (*англ.* *d*-separation) вершин A и B байесовской сети означает, что в случае введения новой информации (свидетельства) e имеет место равенство $P(A\,|\,B,e)=p(A\,|\,e)$. Понятие *d*-разделимости является выражением свойства человеческого мышления, которое состоит в определении конфигурации причинных связей между двумя переменными таким образом, что новая информация об одной переменной не влияет на степень неопределенности другой.

➢ *Маржинализация* означает вычисление суммы вероятностей по реализациях всех переменных, кроме выбранных. Она используется для вычисления вероятностей переменных, которые нас интересуют, на основании погной вероятности:

$$P(A)=\sum_B P(A,B).\qquad (6)$$

➢ *Логический вывод* в сетях Байеса означает вычисление условных вероятностей некоторых переменных на основании имеющейся информации о других переменных.

➢ *Распространение* в БС означает процесс вычисления апостериорных вероятностей для тех узлов сети, которые не наблюдаются, на основании значений наблюдаемых узлов (то есть, свидетельств).

БС называют причинной (каузальной), когда все ее связи носят причинный характер. Можно выделить следующие факторы, определяющие использование причинных моделей в искусственном интеллекте:

– как правило, человек интерпретирует события с точки зрения причина-следствие, что упрощает понимание причинных моделей пользователем;

– идентификация инвариантных причинных связей в конкретной задаче, позволяет спрогнозировать эффекты, возникающие вследствие случайных событий (случайных переменных), и эффекты, обусловленные предопределенными действиями (то есть, действиями, имеющими место внутри рассматриваемой системы или внешними влияниями); при этом под идентификацией инвариантных причинных связей будем понимать определение характера и глубины повторяющихся связей;

– причинность и вероятность тесно связаны между собой, потому что причинность обычно предусматривает существование вероятностных взаимозависимостей, обеспечивающих понимание причинности; фактически, необходимым условием существования причинности есть корреляция;

– аксиоматические свойства БС ($d-$разделение и марковость) соответствуют вероятностным зависимостям и случаям их отсутствия, возникающим в причинной области;

– существуют так называемые канонические вероятностные модели или общепринятые структурные элементы БС (зашумленные И, ИЛИ, МАКС и другие), которые базируются на интерпретации родительских узлов как причин или условий для этого узла, а также на предположении независимости причинных взаимодействий; эти модели снижают число параметров сети, упрощают получение знаний и даже способствуют снижению вычислительных затрат [5];

– причинные БС поддерживают некоторые качественные аспекты принятия решений, которые можно идентифицировать с целью объяснения результатов вывода; некоторые из этих аспектов присущи определенным каноническим моделям, например, генерирование объяснения – явление, типичное для так называемого зашумленного ИЛИ (см. ниже);

– наконец, концепция объяснения результата очень близка понятию причинности; фактически, одним из элементов объяснения результатов тех или иных действий является выяснение влияния зарегистрированных фактов на анализируемую систему.

Некоторые методы объяснения результатов взаимодействия событий и фактов, используемые в БС, разработаны специально для причинных сетей или даже для конкретных канонических моделей, а другие методы являются общими в том смысле, что они не предполагают причинной интерпретации сети.

Остановимся немного подробнее на канонических моделях причинных связей или причинного взаимодействия. Одной из наиболее часто используемых моделей такого типа есть модель дизъюнктивного взаимодействия («зашумленное ИЛИ»). Дизъюнктивное взаимодействие имеет место в случае, когда существует вероятность того, что любое из множества условий может привести к появлению некоторого события и эта вероятность не уменьшается, когда несколько из этих условий возникают одновременно. Например, если есть несколько заболеваний, каждое из которых может вызвать высокую температуру, то при этом только повышается вероятность того, что у пациента, страдающего всеми этими заболеваниями одновременно, будет высокая температура. Более того, если пациент при этом также страдает болезнью, которая (в отдельности) не вызывает высокую температуру, то эта информация не снижает вероятности повышения температуры от других болезней. Такой тип взаимодействия можно аппроксимировать с помощью двух предположений.

(1) Событие A считается ложным ($p(A)=0$), если все условия, рассматриваемые как возможные причины возникновения A, являются ложными. В примере с пациентом это предположение означает, что явно перечисляются основные болезни (причины), которые могут привести к повышению температуры, и все эти причины объединяются вместе под названием «другие причины».

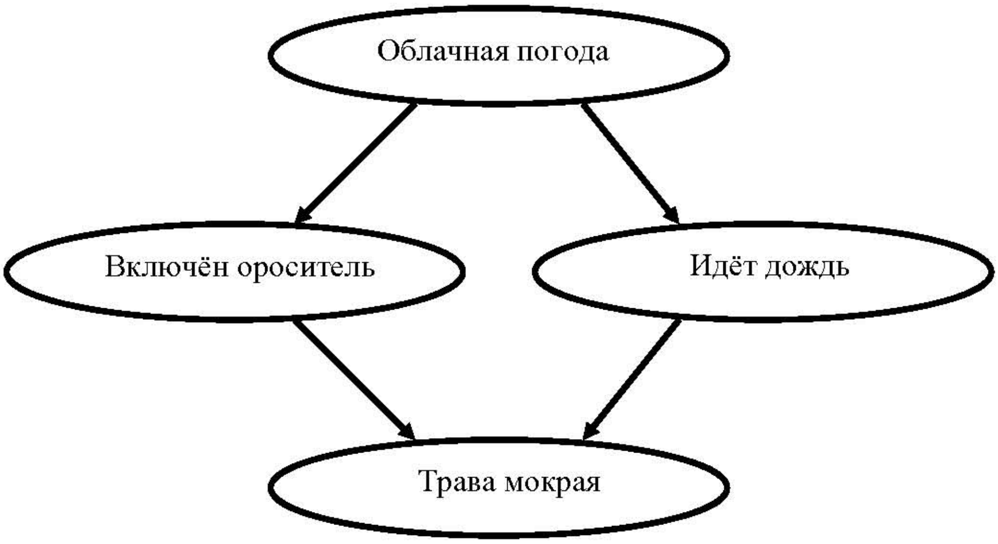
(2) Предполагается, что если событие A является типичным следствием двух условий c_1 и c_2, то механизм запрещения появления события A при условии существования c_1 не зависит от механизма, который может запретить появление события A при условии существования c_2. При этом каждое исключение действует как независимая переменная. Например, система защиты дома может дать сигнал тревоги в результате землетрясения или вторжения вора. В этом случае механизм, который запрещает активацию тревоги при землетрясении, может основываться, например, на слабой чувствительности датчика к вертикальному ускорению. А механизм, действующий в случае вторжения вора, может основываться на учете его искусности и опыте. Эти два механизма можно считать независимыми один от другого. С другой стороны, если возможно исчезновение электричества, то это приведет к запрещению обоих упомянутых механизмов и второе предположение не будет выполняться.

Самой распространенной задачей, которая решается с помощью байесовской сети, есть вероятностный вывод. Например, рассмотрим простую сеть, описывающую функционирование оросителя травы [1].

C	P(S=F)	P(S=T)
F	0,5	0,5
T	0,9	0,1

P(C=F)	P(C=T)
0,5	0,5

C	P(R=F)	P(R=T)
F	0,8	0,2
T	0,2	0,8



S	R	P(W=F)	P(W=T)
F	F	1,0	0
T	F	0,1	0,9
F	T	0,1	0,9

Рис. 1. Байесовская сеть “Дождевая установка”

Предположим, что в результате наблюдения мы получаем информацию, что трава мокрая. Такому состоянию травы есть две причины: идет дождь или включен ороситель. Какое из этих событий имеет более высокую вероятность? Для вычисления апостериорной вероятности каждого из событий можно воспользоваться правилом Байеса (введем обозначения: истина = 1; ложь = 0; W – трава мокрая; S – ороситель; R – дождь). Соответственно, W=1 означает, что трава мокрая; S=1 означает, что ороситель включен; R=1 означает, что идет дождь. Используя правило Байеса, получим:

$$p(S=1|W=1)=\frac{p(S=1,W=1)}{p(W=1)}=\frac{\sum_{c,r} p(C=c,S=1,R=r,W=1)}{p(W=1)}=$$

$$=\frac{0{,}2781}{0{,}6471}=0{,}4298,$$

$$p(R=1|W=1)=\frac{p(R=1,W=1)}{p(W=1)}=\frac{\sum_{c,s} p(C=c,S=s,R=1,W=1)}{p(W=1)}=$$

$$=\frac{0{,}4581}{0{,}6471}=0{,}7079.$$

где

$$P(W=1)=\sum_{c,r,s} P(c=c,S=s,R=r,W=1)=0{,}6471.$$

2 Последовательность построения Байесовской сети

1. Анализ процесса. Сбор данных и экспертных оценок.
2. Формирование базы данных.
3. Генерация топологии сети (узлы и дуги).
4. Определение априорных вероятностей и оптимизация топологии сети.
5. Обучение сети.
6. Использование сети для классификации.
7. Представление результатов пользователю.

3 Типы байесовских сетей

3.1. Дискретные БС – сети, у которых переменные узлов являются дискретными величинами.

Дискретные БС обладает следующими свойствами:

– каждая вершина представляет собой событие, описываемое случайной величиной, которая может иметь несколько состояний;

– все вершины, связанные с “родительскими” определяются таблицей условных вероятностей или функцией условных вероятностей;

– для вершин без “родителей” вероятности её состояний являются безусловными (маргинальными).

Другими словами, в байесовских сетях доверия вершины представляют собой случайные переменные, а дуги – вероятностные зависимости, которые определяются через таблицы условных вероятностей. Таблица условных вероятностей каждой вершины содержит вероятности состояний этой вершины при условии состояний её “родителей” [12]. На рис. 1 приведён пример дискретной БС.

3.2. Динамические БС – сети, у которых значения узлов изменяется со временем.

Динамические БС идеально подходят для моделирования временных процессов. Их преимущество в том что они используют табличное представления условных вероятностей что облегчает представление различных нелинейных явлений [7]. Стоить заметить что термин "временная Байесовская сеть" (temporal Bayesian network) лучше подходит чем "динамическая Байесовская сеть" (dynamic Bayesian network), так как предполагается, что структура модели не изменяется. Также обычно параметры модели не изменяются со временем, однако всегда можно добавить дополнительные скрытые узлы, чтобы описать текущее состояние [1].

Самый простой тип динамической БС - это скрытая модель Маркова (Hidden Markov Model), у которой в каждом слое есть один дискретный скрытый узел и один дискретный или непрерывный наблюдаемый узел. Иллюстрация модели ниже. Круглые вершины обозначают непрерывные узлы, квадратные обозначают дискретные. X – скрытые узлы, а Y – наблюдаемые. Для задания динамической БС, нужно определить начальное распределение $P(X(t))$, топологию внутри слоя $P(X(t+i)|X(t+i-1))$ и межслойную топологию (между двумя слоями) $P(Y(t)|X(t))$ [8].

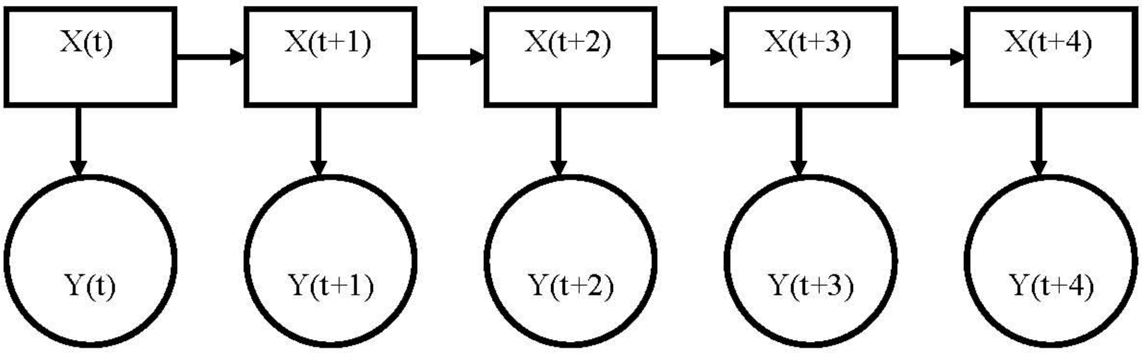


Рис. 2. Скрытая двухслойная модель Маркова, в которой X – скрытые дискретные узлы, а Y – дискретный или непрерывный наблюдаемые

Сети такого вида используют при распознавании речи. В этом случае узлы $Y(t), Y(t+1), Y(t+2), \ldots$ представляют собой фонемами при произношении слов, а узлы $X(t), X(t+1), X(t+2), \ldots$ - это буквы из которых состоит произносимое слово.

Такая модель будет динамической в том смысле, что данная сеть будет представлять собой множество повторяющихся блоков в разные моменты времени [13].

3.3. Непрерывные БС – переменные узлов сети являются непрерывными величинами. Во многих случаях события могут принимать любые состояния из некоторого диапазона. То есть переменная X будет являться непрерывной случайной величиной, пространством возможных состояний которой будет весь диапазон допустимых её значений

$$X = \{ x | a \le x \le b \},$$

содержащий бесконечное множество точек. При этом уже нельзя говорить о вероятности отдельного состояния, так как при бесконечно большом их числе вес каждого будет равен нулю. Поэтому распределение вероятностей для непрерывной случайной величины определяется иначе, чем в дискретном случае и для их описания используются функции распределения вероятностей и плотности распределения вероятностей.

Непрерывные БС используются для моделирования стохастических процессов в пространстве состояний с непрерывным временем [11].

На рис. 3 приведён пример непрерывной БС, в данном случае используется распределение Гаусса. Пусть узел X имеет множество родителей $U = \{U_1, U_2, \ldots, U_n\}$, тогда условное распределение для X задаётся по формуле $f(X | U_i) = N(x; \mu_x + b_i \cdot \mu_i, \sigma_x)$, где коэффициент b_i показывает связь между X и его i-м родителем (его ещё называют весовым коэффициентом) [9, 10]:

$$N(x; \mu_x + b_i \cdot \mu_i, \sigma_x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_x}} \cdot \exp\left(\frac{1}{2} \cdot \frac{(x - (\mu_x + b_i \cdot \mu_i))^2}{\sigma_x} \right),$$

где μ - математическое ожидание, σ - дисперсия, а коэффициент $\frac{1}{\sqrt{2 \cdot \pi \cdot \sigma_x}}$ называется нормирующей константой, которая гарантирует что $\int_x N(x; \mu_x + b_i \cdot x_i, \sigma_x) = 1$.

Связь между переменной X и её родителями $(U_1, U_2, \ldots, U_n)$ можно представить при помощи обычной линейной регрессионной модели:

$$X = b_1 \cdot U_1 + b_2 \cdot U_2 + \ldots + b_n \cdot U_n + Q_x,$$

где Q_x - шумовая компонента, которая может быть записана в виде распределения Гаусса с нулевым математическим ожиданием, а $b_1, b_2, \ldots, b_n$ регрессионные коэффициенты, показывающие связь между переменной X и её предками $U_1, U_2, \ldots, U_n$.

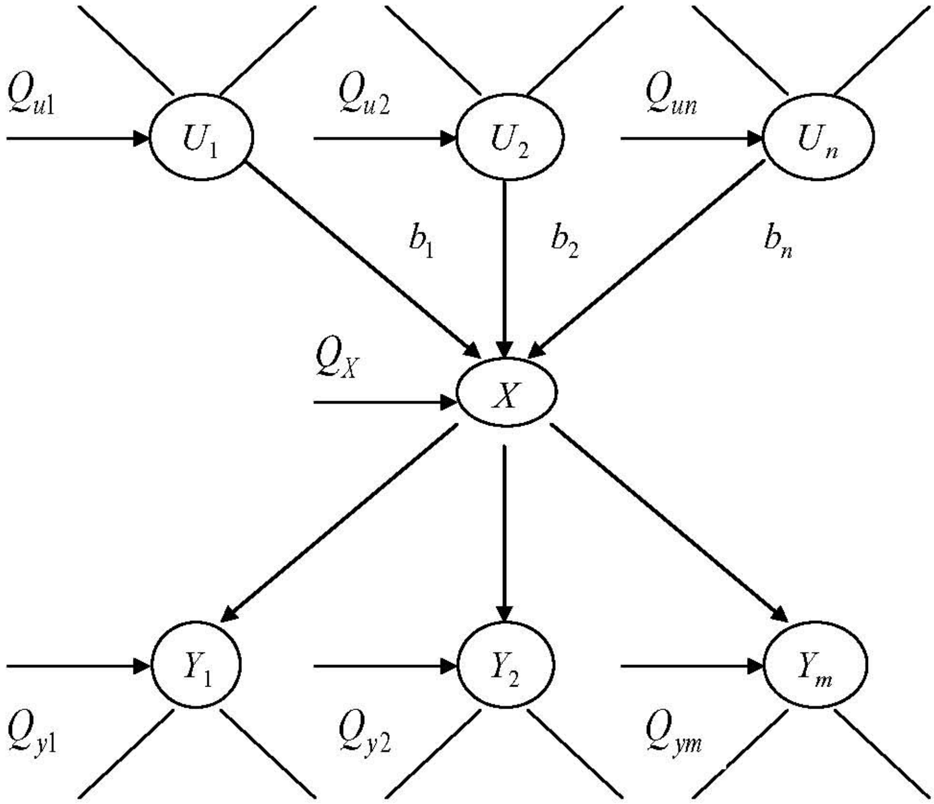


Рис. 3. Пример непрерывной БС. $U_1, U_2, \ldots, U_n$ - предки X, $Y_1, Y_2, \ldots, Y_m$ - дочерние узлы переменной X, $b_1, b_2, \ldots, b_n$ - весовые коэффициенты, $Q_x, Q_{u1}, \ldots, Q_{un}, Q_{y1}, \ldots, Q_{ym}$ - шумовые коэффициенты.

3.4. Гибридные БС – сети, содержащие как узлы с дискретными переменными, так и с непрерывными.

При использовании БС, содержащих как непрерывные, так и дискретные переменные существует ряд ограничений:

1. дискретные переменные не могут иметь непрерывных родителей;

2. непрерывные переменные должны иметь нормальный закон распределения, условный на значениях родителей;

3. распределение непрерывной переменной X с дискретными родителями Y и непрерывными родителями Z является нормальным распределением

$$P(X|Y = y, Z = z) = N(\mu_x(\mu_y, \mu_z), \sqrt{\sigma_x(\sqrt{\sigma_y})}),$$

где μ_x, μ_y, μ_z - математические ожидания, σ_x, σ_y - дисперсии, $\sqrt{\sigma_x}, \sqrt{\sigma_y}$ - среднеквадратические отклонения. μ_x линейно зависит от непрерывных родителей, σ_x вообще не зависит от непрерывных родителей. Однако, оба они (μ_x и σ_x) зависят от дискретных родителей. Это ограничение гарантирует возможность точного вывода.

В качестве примера рассмотрим гибридную БС, показанную на рис. 4, которая позволяет оценивать суммарные производственные затраты в зависимости от использования и загрузки трёх групп оборудования (например, трех станков) [12]. Известно что в состав суммарных производственных затрат (без учёта зарплаты и начислений) входят:

– прямые производственные затраты на каждую группу оборудования за исследуемый календарный период, которые зависят как от количества используемых групп оборудования, так и от времени работы каждой из групп в течение исследуемого периода времени, т.е. от степени загрузки каждой из групп;

– расходы на амортизацию каждой из групп оборудования, зависящие как от её балансовой стоимости, так и установленных норм амортизации;

– арендная плата за участок при каждой из групп оборудования, используемый для складирования сырья и продукции, которая зависит как от площади участка, так и от ставок арендной платы.

Вершина «Загрузка оборудования» соответствует дискретному событию, которое характеризуется тремя возможными состояниями. Вероятность пребывания в каждом из них определяется степенью загрузки каждой из групп оборудования, при условии, что суммарная загрузка всего оборудования равна единице. Будем считать, что все группы оборудования загружены равномерно. На самом деле возможны и любые другие исходные распределения вероятностей, учитывающие различные варианты загрузки оборудования.

Пусть ставка аренды 1 га земли в среднем составляет 2500 у.е. и колеблется в пределах ±10%, то есть принимает значения 2500 ± 250 у.е. Следовательно этой вершине соответствуют параметры $\mu = 2500$ и $\sigma = 62500 = (250)^2$. А норма амортизации может находиться в пределах 5 - 10 % от балансовой стоимости, то есть принимать значения 7,5 ± 2,5% (или 0,075 ± 0,025 относительных единиц). То есть этой вершине соответствуют параметры $\mu = 0{,}075$ и $\sigma = 0{,}000625 = (0{,}025)^2$.

Что касается вершины «Производственные затраты», то она характеризуется случайной переменной, условно нормальной на значениях родителей (то есть на значениях трёх других вершин нашего примера). Следует отметить, что в общем случае распределение вероятностей для вершин, аналогичных «Производственные затраты» является не просто нормальным, а смешанным нормальным распределением. То есть представляет собой весовую сумму распределений, для каждого из которых должен быть задан список его параметров:

– математические ожидания и дисперсии для распределений, описывающих степень влияния дискретных родителей;

– весовые коэффициенты, учитывающие степень влияния на математическое ожидание непрерывных родителей.

Предположим, что в нашем примере выполняются условия:

– балансовая стоимость каждой из пилорам составляет 50000, 40000 и 30000 у.е. ,

– площадь арендуемых участков, закрепляемая за ними равна 0,6; 0,5 и 0,4 га,

– а оценка прямых затрат на поддержание нормальной работы каждой из пилорам в среднем составляет 3000, 3200 и 3500 у.е. и получена с 5% точностью.

То есть, степень влияния родительских вершин на «Производственные затраты» можно представить в виде таблицы 1.

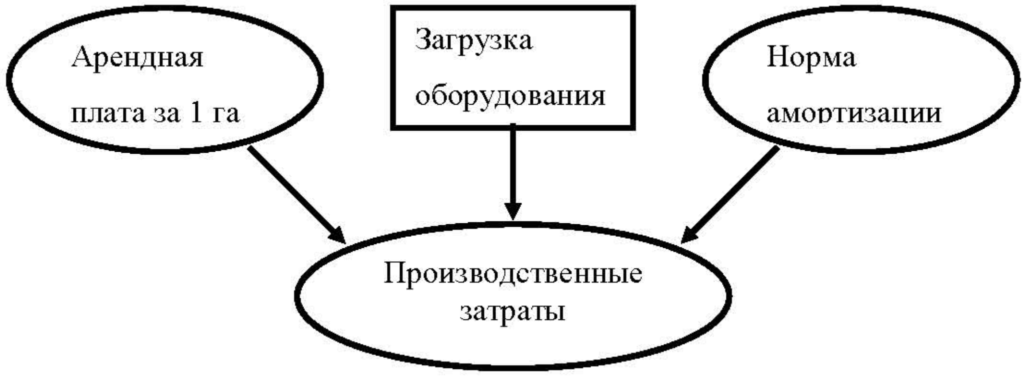
Таблица 1. Параметры, определяющие распределение вероятностей для вершины «Производственные затраты»

Загрузка оборудования	Станок 1	Станок 2	Станок 3
μ	3000	3200	3500
σ	22 500 = 150*150	25 600 = 160*160	30 625 = 175*175
Норма амортизации	50 000	40 000	30 000
Ставка аренды	0,6	0,5	0,4

	Арендная плата
μ	2500
σ	62500= 250*250

P(Загрузка оборудования) = Станок 1	0,333
P(Загрузка оборудования) = Станок 2	0,333
P(Загрузка оборудования) = Станок 3	0,333

	Норма амортизации
μ	0,075
σ	0,000625= 0,025*0,025



	Производственные затраты
μ	7483,33
σ	1459505,06=1208,1*1208,1

Рис. 4. Пример гибридной БС с непрерывными и дискретными событиями. Квадратные вершины соответствуют дискретным событиям, а овальные - непрерывным событиям (гауссовым переменным)

Логический вывод в таких БСД заключается в распространении вероятностей и параметров гауссовых законов распределения по всей сети в зависимости от полученных свидетельств. В основе процесса логического вывода лежат довольно сложные математические алгоритмы, которые мы рассмотрим на простейшей двухуровневой сети для случая прямого распространения распределения вероятностей.

Пусть независимые дискретные случайные величины $X_1,\ldots,X_S$ и непрерывные случайные величины $Z_1,\ldots,Z_r$ оказывают влияние на результирующую случайную величину Y, как показано на рис. 5.

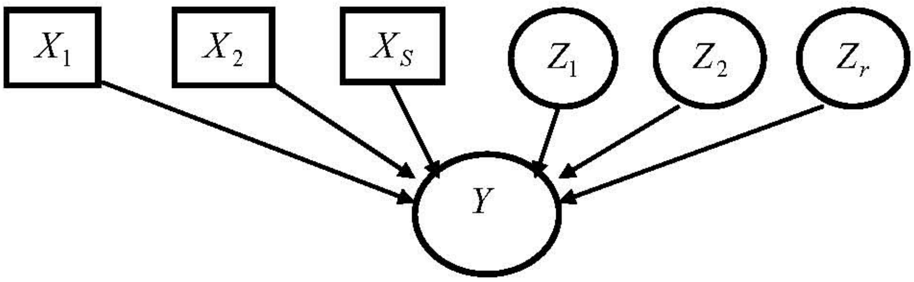


Рис. 5. Пример двухуровневой гибридной БС с непрерывными и дискретными переменными. Квадратные вершины соответствуют дискретным событиям, а овальные - непрерывным событиям.

Каждая из дискретных случайных величин $X_j (j=1,\dots,s)$ имеет своими исходами значения $X_{i_i} (i=1,\dots,n_j)$ с вероятностями P_{ij}, для которых $\sum_{i=1}^{n_j} P_{ij} = 1$. Совместное влияние дискретных случайных величин на Y характеризуется математическим ожиданием $(\mu_{i1},\dots,\mu_{is})$ и дисперсиями $(\sigma_{i1},\dots,\sigma_{is})$. Каждая из непрерывных случайных величин Z_l имеет непрерывное нормальное распределение с параметрами (μ_l,σ_l), где $l=1,\dots,r$. Совместное влияние непрерывной случайной величины Z_l и исходов дискретных величин на результирующую случайную величину Y характеризуется весовыми коэффициентами $k_{l,i1,\dots is}$ для $l=1,\dots,r$.

Тогда характеристики результирующей величины Y могут быть вычислены по следующим выражениям:

$$\mu = \sum_{i1}^{n1} \cdots \sum_{is}^{ns} p_{1,i1} \cdots p_{s,is} \left(\mu_{i1,\dots,is} + \sum_{l=1}^{r} K_{l,i1,\dots,is} \cdot \mu_l \right),$$

$$\sigma = \sum_{i1}^{n1} \cdots \sum_{is}^{ns} p_{1,i1} \cdots p_{s,is} \cdot \left(\left(\mu_{i1,\dots,is} + \sum_{l=1}^{r} K_{l,i1,\dots,is} \right)^2 + \sigma_{i1,\dots,is} + \sum_{l=1}^{r} K_{l,i1,\dots,is}^2 \cdot \sigma_l \right) - \mu^2.$$

Применительно к нашему примеру, содержащему две исходные непрерывные ($r=2$) переменные и одну дискретную ($s=1$) переменную, имеющую три исхода ($n1=3$), числовые характеристики случайной переменной «Производственные затраты» будут равны:

$$\mu = \sum_{i=1}^{3} p_i \left(\mu_i + \sum_{l=1}^{2} k_{li} \cdot \mu_i \right) =$$
$$= 0{,}333 \cdot (3000 + 50000 \cdot 0{,}075 + 0{,}6 \cdot 2500) +$$
$$+ 0{,}333 \cdot (3200 + 40000 \cdot 0{,}075 + 0{,}5 \cdot 2500) +$$
$$+ 0{,}333 \cdot (3500 + 3000 \cdot 0{,}075 + 0{,}4 \cdot 2500) =$$
$$= 0{,}333 \cdot (8250 + 7450 + 6750) = 7483{,}33$$

$$\sigma = \sum_{i=1}^{3} p_i \left(\left(\mu_i + \sum_{l=1}^{2} k_{li} \cdot \mu_l \right)^2 + \sigma_i + \sum_{l=1}^{r} k_{li}^2 \cdot \sigma_l \right) - \mu^2 = 1459505{,}6,$$

$$\sqrt{\sigma} = 1208{,}1.$$

4 Обучение сети

Для описания Байесовской сети необходимо определить топологию графа и параметры каждого узла. Эту информацию мы можем получить из обучаемых данных, но получение правильной топологии сети является более сложной задачей, чем получение параметров узлов [14, 15, 16, 17]. Особого подхода требует случай, когда некоторые из узлов скрыты, или мы имеем дело с некорректными или недостаточными данными. Поэтому существует 4 случая, которые приведены в таблице 1.

Таблица 1. Четыре случая обучения сети

Структура	Наблюдение	Метод
Известна	Полное	Максимальная Оценка Правдоподобия
Известна	Частичное	максимизации математического ожидания или жадный метод поиска экстремума
Незвестна	Полное	Поиск в пространстве моделей
Незвестна	Частичное	Структурный алгоритм максимизации математического ожидания или сжатия границ

4.1. Известная структура, полная наблюдаемость

В этом случае вычисляются значения параметров каждого условного вероятностного распределения, которые максимизируют правдоподобность обучающих данных. Нормализованное логарифмическое уравнение правдоподобия имеет вид:

$$L = \frac{1}{N} \cdot \sum_{i=1}^{m} \sum_{l=1}^{s} \log\left(P\left(X_i \middle| Pa(X_i), D_l\right)\right),$$

где $Pa(X_i)$ обозначает набор родителей X_i, D - обучающее множество, а D_l - это l-й элемент обучающего множества, N - количество событий в обучающем множестве.

Для дискретного случая показанного на рис. 1, оценка узла W вычисляется следующим образом. Пусть у нас есть данные для обучения, то есть информация о том сколько раз трава была влажная, а при причиной этого было то что шёл дождь и был включен разбрызгиватель $N(W=1, S=1, R=1)$, сколько раз трава была влажная потому что шёл дождь, а разбрызгиватель был выключен $N(W=1, S=0, R=1)$ и так далее. Используя эти данные максимальная оценка правдоподобия узла W вычисляться как:

$$P(W=w|S=s, R=r) = \frac{N(W=w, S=s, R=r)}{N(S=s, R=r)} = \\ = \frac{N(W=w, S=s, R=r)}{N(W=0, S=s, R=r) + N(W=1, S=s, R=s)}.$$

Если вершины описаны функцией Гаусса, то можно посчитать выборочное среднее значение и дисперсию, а затем при помощи линейной регрессии посчитать матрицу весов. Для других видов распределений, используются более сложные процедуры.

4.2. Известная структура, частичная наблюдаемость

Когда некоторые из узлов скрыты, то можно применить алгоритм максимизации математического ожидания (ММО), для нахождения локальной оптимальной оценки максимального правдоподобия (ОМП) параметров. Основная идея алгоритма ММО состоит в том что, если бы мы знали значения всех узлов, обучение (на шаге М) было бы простым, поскольку знакомы с предыдущими. Так на шаге Е, мы вычисляем ожидаемые значения всех узлов, использующих алгоритм вывода, и затем используем эти значения как если бы они были получены из наблюдений.

Для узла W из примера приведённого на рис. 1, выполняется замена наблюдаемого события ожидаемым количеством выполнения события:

$$P(W=w|S=s, R=r) = \frac{EN(W=w, S=s, R=r)}{EN(S=s, R=r)}$$

где $EN(x)$ - ожидаемое количество выпадения события x в период обучения, при этом учитывается текущая оценка параметров. Параметр $EN(x)$ можно считать по формуле:

$$EN(x) = E\sum_k I(x|D(k)) = \sum_k P(x|D(k))$$

где $I(x|D(k))$ - функция-индикатор, которая принимает значение 1, если в обучающем процессе $D(k)$ происходит событие x, в противном случае функция принимает значение 0.

Учитывая ожидаемое количество выполнения события, мы максимизируем параметры, а затем повторно вычисляем ожидаемое количество выполнения события и так далее. Этот итеративный метод сходиться к локальному максимуму значения вероятности.

4.3. Неизвестная структура, полная наблюдаемость

Наиболее вероятной моделью в данном случае является полный граф, потому что в этом случае будет задействовано наибольшее количество параметров, следовательно такая модель будет больше всего соответствовать данным.

Формула Байеса имеет вид:

$$P(G|D) = \frac{P(D|G) \cdot P(G)}{P(D)},$$

где G направленным нециклический граф, соответствующий случайным переменным, а $D = \{x^1, \ldots, x^N\}$ множество данных, прологарифмируем её:

$$\log(P(G|D)) = \log(P(D|G)) + \log(P(G)) + (-\log(P(D))).$$

В полученном выражении слагаемое $-\log(P(D))$ играет роль штрафующей компоненты за чрезмерно сложные модели. Слагаемо $P(D|G)$ также может выступать в качестве штрафующей компоненты за чрезмерно сложные модели (этот случай известен как бритва Оккама, более подробно об этом в [18]).

Для выполнения точных расчётов связанных с выбором модели требуется вычислить $P(D) = \sum_G P(D,G)$, что является задачей экспоненциальной сложности. Вместо этого можно использовать БИК (Байесовский информационный критерий), который определяется как:

$$\log(P(G|D)) \approx \log(P(D|G,\hat{\theta}_G)) - \frac{\log(N)}{2} \cdot \dim(G),$$

где N – количество моделей, $\dim(G)$ – размер модели, $\hat{\theta}_G$ – максимально правдоподобная оценка параметров, слагаемое $-\frac{\log(N)}{2} \cdot \dim(G)$ играет роль штрафующей компоненты за чрезмерно сложные модели [16].

Следующим шагом после выбора структуры является обучение структуры, так что бы направленный нециклический граф лучше всего удовлетворял данным. Эта задача является НП-трудной, то есть задачей с нелинейной полиномиальной оценкой числа итераций. Поэтому обычно используют локальные алгоритмы поиска (например, жадный метод поиска экстремума (greedy hill climbing method)) или метод ветвей и границ, для поиска в пространстве графов.

Так же для построения байесовской сети по записям из базы данных можно использовать алгоритм К2, где D - база вариантов; Z - множество переменных соответствующих D; B_{si}, B_{sj} - две структуры байесовской сети, содержащие переменные из Z. Подсчитав попарно соотношения вида

$$\frac{P(B_{S_i}|D)}{P(B_{S_j}|D)} = \frac{\dfrac{P(B_{S_i},D)}{P(D)}}{\dfrac{P(B_{S_j},D)}{P(D)}} = \frac{P(B_{S_i},D)}{P(B_{S_j},D)}$$

между всеми структурами байесовской сети, можно ранжировать структуры относительно их апостериорных (позднейших) вероятностей. Для обучения сети используют следующую функцию:

$$g(i,\pi_i) = \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!,$$

где r_i - номер состояния x_i в Z, т.е. x_i имеет r_i возможных значений: $(v_{i1},\ldots,v_{ir_i})$, а N_{ijk} связывает множество предков π_i для x_i с множеством данных D. При этом предок формально определяется как $pred(x_i) = \{x_1,\ldots,x_{i-1}\}$. Алгоритм возвращает множество вершин, предшествующих вершине x_i, в соответствии с расположением вершин. Входными данными являются множество вершин, расположение вершин, ограничение на количество предков для вершины (обозначим как u) и база данных содержащая m случаев. На выходе получаем множество предков для каждой вершины [21].

4.4 Неизвестная структура, частичная наблюдаемость

Это самый сложный случай, когда структура неизвестна и есть скрытые переменные и некорректные данные. В этом случае используют структурный алгоритм максимизации математического ожидания (СММО) [20] или алгоритм сжатия границ (Bound and Collapse) [19].

Алгоритм СММО соединяет в себе стандартный алгоритм ММО, который оптимизирует параметры, со структурным поиском модели отбора. Этот алгоритм обучает сети, основываясь на штрафных вероятностных значениях, которые включают значения, полученные с помощью байесовского информационного критерия, принципа минимальной длины описания, а также значения других критериев.

Метод сжатия границ (СГ) моделируют отсутствие данных, предполагая что вероятность отсутствие данных находится в интервале от 0 до 1. То есть производится вычисление этого интервала отсутствия данных, по имеющейся информации. После этого производится сжатие границ интервала в точку посредством использования выпуклой комбинации точек экстремумов, основываясь на информации о неполных данных.

Выводы

Байесовские сети – перспективный подход к моделированию процессов с неопределенностями различной природы. Они могут быть использованы как для моделирования статических, так и динамических процессов. В данной работе было дано формальное определение понятия байесовская сеть, рассмотрены общие принципы и последовательность построения байесовских сетей, приведены типы байесовских сетей, а также рассмотрена задача обучения байесовской сети. Для каждого типа байесовской сети приведены практические примеры использования. Даны краткие описания методов, используемых для обучения байесовской сети в различных ситуациях. В дальнейшем предполагается применение БС для решения конкретных задач с использованием оригинальных методов обучения и вывода.

Литература

1. Murphy K. A Brief introduction to graphical models and Bayesian networks / http://www.cs.Berkeley.edu, 2001. – 19 p.
2. Pearl J. Causality. – Cambridge: Cambridge University, 2000. – 384 p.
3. Druzdzel M.J., Gaag L.C. Building probabilistic networks: where do the numbers come from? // IEEE Transactions on Knowledge and Data Engineering, v.12, No.4, 2000, 481-486 p.
4. Heckerman D., Breeze J.S. Causal independence for probability assessment and inference using Bayesian networks / Technical report MSR-TR-94-08. – Microsoft Research, Redmond, WA (USA), 1995. – 14 p.
5. Lauritzen S.L., Spiegelhalter D.J. - "Local computations with probabilities on graphical structures and their application to expert systems" in Journal Royal Statistics Society B, 50(2), 1988. 157-194 p.
6. Buntine V. L. Learning with Bayesian Models / Report 94-04-13, NASA Ames Research Center, 1994. – 78 pp.
7. Zweig G.G. Speech Recognition with Dynamic Bayesian Networks / university of California, Berkeley, http://www.cse.msu.edu, 1998. – 169 p.
8. Haipeng Guo. Dynamic Bayesian Networks / http://www.kddresearch.org, 2002. – 21p.
9. Hautaniemi S.K. Target Identification with Bayesian Networks. Master of Science thesis. / http://www.cs.tut.fi/~samba/Publications, 2000 - 99 p.
10. Hautaniemi S.K., Petri T. Korpisaari and Jukka P.P. Saarinen. Target Identification with Dynamic Hybrid Bayesian Networks. 2000 – 11 p.
11. Nodelman U., Christian R. Shelton, and Daphne Koller. Learning Continuous Time Bayesian Networks. // Proceedings of the Nineteenth International Conference on Uncertainty in Artificial Intelligence, 2003, 451-458 p.
12. Хабаров С.П. Экспертные системы. Конспект лекций. / http://firm.trade.spb.ru/serp, 2001 – 7 стр.
13. Buntine W. L. A Guide to the Literature on Learning Probabilistic Networks from Data. IEEE Transactions on Knowledge and Data Engineering, volume 8 , issue 2, April 1996, 195 – 210 p
14. Heckerman D. A tutorial on learning with Bayesian networks / Technical report MSR-TR-95-06. – Microsoft Research, Redmond, WA (USA), 1995. – 57 p.

15. Ferat S. A Bayesian Network Approach to the Self-organization and Learning in Intelligent Agents / Dissertation submitted to the Faculty of Virginia Polytechnic and State University in fulfillment of the requirements for the degree of Doctor of Philosophy in Electrical and Computer Engineering, 2000 – 251 p.
16. Murphy K. P. An introduction to graphical model. http://www.ai.mit.edu. 2001 – 19 p.
17. Murphy K. and Mian S. Modelling Gene Expression Data using Dynamic Bayesian Networks. Technical report, Berkeley, CA. / http://citeseer.ist.psu.edu. 1999 - 12 p.
18. MacKay D. J. C. Probable networks and plausible prediction – a review of practical Bayesian methods for supervised neural networks. Journal "Network Computation in Neural Systems" № 6. 1995, 469-505 p.
19. Sebastiani P. and Ramoni M., "Bayesian inference with missing data using bound and collapse," Technical Report KMi-TR-58 / http://chip.tch.harvard.edu. 1997 – 21 p.
20. Friedman N. The Bayesian Structural EM Algorithm. Fourteenth conference on Uncertainty in Artificial Intelligence (UAI) / http://www.cs.huji.ac.il. 1998 – 10 p.
21. Guo H. Learning Bayes Networks from Data. http://www.kddresearch.org. 2000 - 17.

Abstract

This analytical literature review discusses different methods under the general rubric of learning Bayesian networks from data. The basic concepts of Bayesian networks and their learning methods are introduced and reviewed. The methods are discussed for learning parameters of a probabilistic network, for learning the structure, and for learning hidden variables. Basic definitions and key concepts with appropriate illustrative examples are presented.

Реферат

Рассмотрены различные методы обучения байесовских сетей на основе данных. Рассмотрены основные методы обучения, а также методы обучения параметров, структуры сети и скрытых параметров. Приведены основные определения и соответствующие иллюстративные примеры.

Реферат

Розглянуто методи навчання байєсових мереж на основі даних. Розглянуто основні методи навчання, а також методи навчання параметрів, структури мережі і латентних параметрів. Наведені основні визначення та відповідні ілюстративні приклади.

Сведения про авторов

Бидюк Петр Иванович, доктор технических наук, профессор Института прикладного системного анализа Национального технического университета Украины «Киевский политехнический институт».

E-mail: peterb@mmsa.ntu-kpi.kiev.ua

Терентьев Александр Николаевич, аспирант Института прикладного системного анализа Национального технического университета Украины «Киевский политехнический институт».

E-mail: paladin@astral.ntu-kpi.kiev.ua

Гасанов Айдын Сардар оглы, кандидат технических наук, старший научный сотрудник МНУЦ.

E-mail:

Constructing and Methods of Learning of Bayesian Networks
Побудова і методи навчання байєсових мереж

Ключевые слова
Графические модели, Байесовские сети, алгоритм построения, методы обучения, особенности структур, примеры построения сетей

Графічні моделі, мережі Байєса, алгоритм побудови, методи навчання, особливості структур, приклади побудови мереж

Graphical models, Bayesian nets, constructing algorithms, methods of learning, specific features of structures, examples of nets.