

УДК 519.766.4

## Методика построения сетевых вероятностных моделей

О.А. Кожуховская

Черкасский государственный технологический университет  
Черкассы, Украина, [olga-kozuhovska@mail.ru](mailto:olga-kozuhovska@mail.ru)

Предложена практическая методика построения графических вероятностных моделей для решения задач прогнозирования, классификации и диагностики. Методика охватывает все этапы построения модели и ее использования.

Ключевые слова: задачи прогнозирования, классификации и диагностики, сетевые вероятностные модели, методика построения.

**Введение.** При решении многих практических задач моделирования процессов различной природы, прогнозирования и поддержки принятия решений возникает необходимость оценивания гипотез, в отношении которых нет полной информации (достаточной) информации. Достаточно часто непросто вычислить приемлемые по качеству оценки переменных и параметров, но, несмотря на наличие неопределенностей, удается найти приемлемые рациональные решения. Компьютерные системы поддержки принятия решений, которые широко используются в различных отраслях деятельности, также должны содержать средства учета неопределенностей при построении моделей исследуемых процессов и формировании альтернатив. Классическими примерами задач с неопределенностями являются техническая и медицинская диагностика, анализ и менеджмент рисков в технике, экономике, финансах, прогнозировании и управлении в условиях неполной информации и т. п.

**Постановка задачи.** Цель работы состоит в анализе проблем, связанных с построением вероятностных моделей в форме байесовских сетей (БС), и разработке практической методики построения БС в условиях полных и неполных наблюдений, которые используются для обучения структуры и параметров модели. Полученные структуры БС должны быть пригодными для решения задач прогнозирования, классификации и диагностики на основе соответствующих статистических данных и экспертных оценок.

Формально, сеть Байеса представляет собой направленный ациклический граф (НАГ) **G** на множестве переменных

$X_1, X_2, \dots, X_n$ , между которыми существуют некоторые причинно-следственные связи. Каждой переменной соответствует вершина графа, а направленные дуги, соединяющие вершины, указывают на существующие зависимости между переменными. Дочерние узловые переменные описывают таблицами условных распределений вероятностей состояний этих переменных при условии, что родительские узлы принимают определенные значения. Совместное распределение вероятностей состояний переменных НАГ определяется выражением:

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n | \mathbf{G}) &= \\ &= P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \mathbf{x}_{pa(i)}, \mathbf{G}), \end{aligned} \quad (1)$$

где  $\mathbf{x}_{pa(i)}$  – вектор непосредственных родительских переменных (узлов) для  $X_i$ ;  $P(x_1, x_2, \dots, x_n | \mathbf{G})$  – вероятность появления конкретной комбинации значений  $x_1, x_2, \dots, x_n$  для множества переменных  $X_1, X_2, \dots, X_n$ . БС структурируются локально таким образом, что каждый узел взаимодействует только со своими родительскими узлами. Условные распределения вероятностей для дискретных переменных представляются множеством соответствующих (многомерных) таблиц с параметрами

$$\Theta = \{\theta_1, \theta_2, \dots, \theta_n\} = \{\theta_{ik}(j)|_{k=1}^{r_i}\}_{j=1}^{q_i},$$

где  $i = 1, \dots, n$  – номера переменных  $X_i \in \mathbf{X}$ ;  $k = 1, \dots, r_i$  – индекс, который указывает на значения переменной  $X_i$ ;  $j = 1, \dots, q_i$  – индекс, указывающий на множество допустимых комбинаций значений родительских переменных для  $X_i$  ( $x_{pa(i)}$ ). Теперь уравнение (1) можно представить в виде:

$$P(\mathbf{x} | \Theta, \mathbf{G}) = \prod_{i=1}^n P(x_i | x_{pa(i)}, \Theta_i, \mathbf{G}) = \prod_{i=1}^n \theta_{ik}(j). \quad (2)$$

Рассмотрим задачу построения байесовской сетевой модели (2) на основе выборки данных мощностью  $N$  значений. Обозначим через

$\mathbf{D} = \{ \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)} \}$  множество векторов данных,

сформированных из значений состояний  $x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}$

переменных  $X_1, X_2, \dots, X_n$ . При этом возможны такие случаи: (1)

– структура БС известна, а необходимо оценить только ее параметры (значения таблиц условных вероятностей); (2) – необходимо оценить структуру и параметры сети. Возможны особые случаи построения модели, когда некоторые вершины скрыты или когда данные неполные. Поэтому в общем случае рассматривают четыре случая обучения сети, которые приведены в табл. 1.

Таблица 1

Четыре случая обучения байесовских сетей

Структура	Наблюдения	Метод
Известна	Полные	Метод максимального правдоподобия
Известна	Неполные	Градиентные методы, EM-алгоритм (максимизация математического ожидания), применение выборки Гиббса
Неизвестна	Полные	Поиск в пространстве моделей
Неизвестна	Неполные	Структурный EM-алгоритм, алгоритм сжатия границ

В случае наличия полных наблюдений для оценивания параметров сети можно воспользоваться методом максимального апостериорного оценивания (МАО). Для дискретных переменных такие оценки представляют собой относительные частоты появления каждого значения для каждой переменной при известной конфигурации родительских узлов. Для реализации процедуры оценивания за методом МАО не обходимо знать (выбрать) априорное распределение для параметров модели. Для этой цели часто используют распределение Дирихле, сопряженное по отношению к многомерному распределению функции правдоподобия.

В случае, если структура сети неизвестна, необходимо оценить сначала структуру графа  $\mathbf{G}$ . Это включает в себя создание спецификаций относительно условной независимости между

переменными модели и параметрами  $\Theta$ . Оценивание структуры сети на основе статистических данных можно выполнить двумя путями: с помощью оптимизационных методов с учетом ограничений и с помощью поисковых процедур на основе скоринговых функций. Методы, которые основываются на ограничениях, достаточно эффективны, но они не имеют практически необходимой робастности. То есть, полученные в результате их применения структуры очень чувствительны к ошибкам статистического тестирования на условную независимость. Таким образом, в большинстве случаев решения практических задач используют поисковые алгоритмы на основе скоринговых функций. Существующие процедуры эвристического поиска дают возможность найти лучшие структуры сетей и связанные с ними распределения вероятностей в пространстве всех возможных структур БС. Среди скоринговых функций популярна байесовская, которая основывается на апостериорной вероятности графа  $G$ ; функция на основе аппроксимации апостериорных распределений вероятностей с использованием информационного критерия Байеса; скоринговые функции на основе описания минимальной длины (ОМД) и информационно-геометрического критерия (Info-geo).

Предлагаемая методика построения сети на основе статистических данных состоит из таких шагов: (1) – постановка задачи исследования; (2) – анализ исследуемого процесса (объекта) и выбор множества переменных для его описания, анализ глубины взаимосвязей между переменными; (3) – редукция размерности задачи моделирования; (4) – масштабирование и дискретизация значений переменных; (5) – определение семантических ограничений; (6) оценивание сетевых моделей-кандидатов; (7) – анализ качества моделей и выбор лучшей из них для решения поставленной задачи исследования.

**Заключение.** Выполнен анализ проблем, связанных с построением вероятностных моделей в форме байесовских сетей. Предложена практическая методика построения БС, пригодных для решения задач прогнозирования, классификации, диагностики и т. п. Методика охватывает все этапы построения модели и ее дальнейшее использование. Значительное внимание уделено обучению структуры модели на основе статистических данных и экспертных оценок. Показано, что существует множество критериев оптимизации структуры (скоринговые функции) и алгоритмов поиска экстремума, которые обеспечивают возможность получения альтернативных структур-кандидатов и выбрать лучшую из них для практического использования.

## **Technique of creation of network probabilistic models**

O. A. Kozhukhovskaya

Cherkassk state technological university  
Cherkassy, Ukraine, olga-kozuhovska@mail.ru

The practical technique of creation of graphic probabilistic models for the solution of problems of forecasting, classification and diagnostics is offered. The technique covers all stages of creation of model and its application.

Keywords: problems of forecasting, classification and diagnostics, network probabilistic models, construction technique.