Marco Scutari Bayesian Network Models for Incomplete and Dynamic Data

Statistica Neerlandica, 2020, 74 (3), p. 397-419 [arXiv:1906.06513](https://arxiv.org/abs/1906.06513)

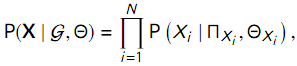
“expert systems” were conceived as the combination of a formal representation of domain-speciﬁc knowledge gathered from subject matter experts and an “inference engine” that could use that representation to answer arbitrary queries. BNs (Bayesian Networks) can represent complex phenomena in a modular way due the conditional independence assumptions they encode, allowing them to scale beyond trivial problems and to be used to develop efﬁcient inference algorithms.

«Экспертные системы» задуманы как комбинация формального представления предметно-ориентированных знаний, собранных у экспертов в предметной области, и «механизма вывода», который мог использовать это представление для ответа на произвольные запросы. BN могут представлять сложные явления в модульном виде из-за предположений условной независимости, которые они кодируют, что позволяет им масштабироваться за пределы тривиальных проблем и использоваться для разработки эффективных алгоритмов вывода.

2 | BACKGROUND AND NOTATION

BNs are a class of graphical models in which the nodes of a directed acyclic graph (DAG) represent a set X of random variables describing some quantities of interest. The arcs connecting those nodes express direct dependence relationships, with graphical separation in implying conditional independence in probability. As a result, induces the factorization

BN - это класс графических моделей, в которых узлы ориентированного ациклического графа (DAG) представляют собой набор X случайных величин, описывающих некоторые представляющие интерес величины. Дуги, соединяющие эти узлы, выражают отношения прямой зависимости с графическим разделением, подразумевающим условную независимость по вероятности. В результате получается разложение

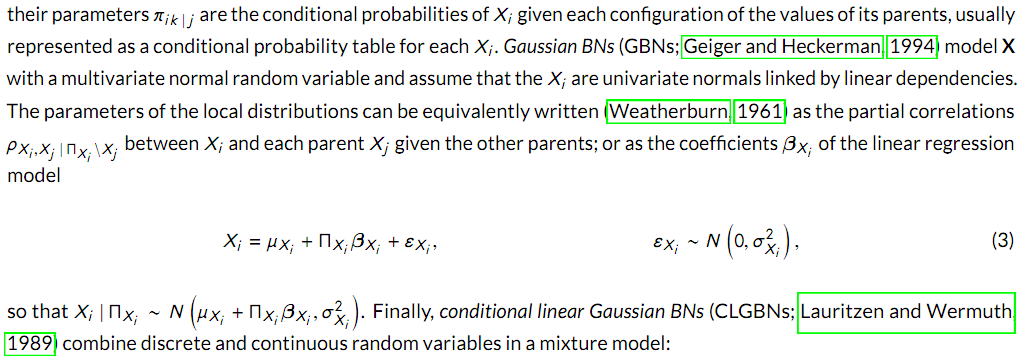


in which the joint probability distribution of X (with parameters Θ) decomposes into one local distribution for each Xi (with parameters , X ) conditional on its parents ПXi. Assuming G is sparse 1, BNs provide a compact representation of both low- and high-dimensional probability distributions.

в котором совместное распределение вероятностей X (с параметрами Θ) распадается на одно локальное распределение для каждого Xi (с параметрами X), обусловленное его родительскими элементами ПXi. Если принять значение G sparse1 (разреженный, редкий), BN обеспечивает компактное представление распределений вероятностей как малой, так и высокой размерности.

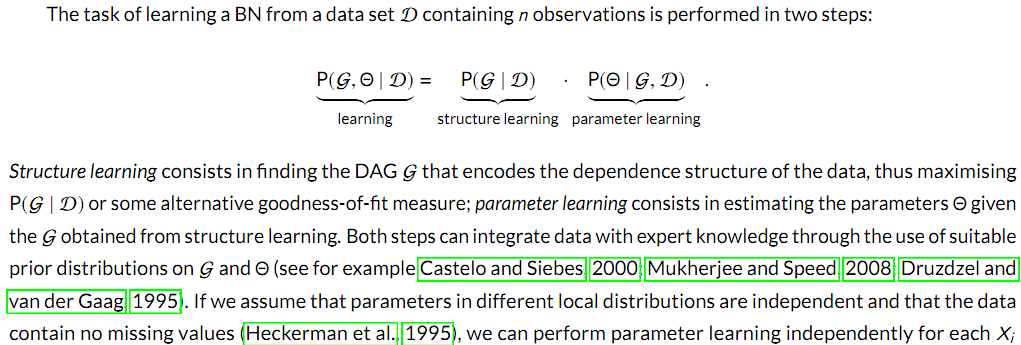
BNs are also very ﬂexible in terms of distributional assumptions; but while in principle we could choose any probability distribution for X, the literature has mostly focused on three cases for analytical and computational reasons. Discrete BNs (Heckerman et al., 1995) assume that both X and the Xi are multinomial random variables. Local distributions take the form 

BN также очень гибки с точки зрения предположений о распределении; но пока в принципе можно было выбрать распределения вероятностей X, литература в основном сосредоточена на трех случаях по аналитическим и вычислительным причинам. Дискретные BN (Heckerman et al., 1995) предполагают, что как X, так Xi и являются полиномиальными случайными величинами. Местные распределения принимают форму



их параметры - это условные вероятности данной каждой конфигурации значений ее родителей, обычно представленные в виде таблицы условных вероятностей для каждой. Гауссовские BN (GBNs; Geiger and Heckerman, 1994) моделируют X многомерной нормальной случайной величиной и предполагают, что это одномерные нормальные величины, связанные линейными зависимостями. Параметры локального распределения могут быть эквивалентно записаны (Weatherburn, 1961) как частичные корреляции между каждым родителем и другими родителями; или как коэффициенты модели линейной регрессии (3).

Наконец, условные линейные Гауссовские BN объединяют дискретные и непрер. переменные в смешанной модели.



Структурное обучение состоит в нахождении графа DAG, который кодирует структуру зависимости данных, таким образом максимизируя P или некоторую альтернативную меру соответствия; Параметрическое обучение состоит в оценке параметров Θ с учетом G, полученного в результате структурного обучения. Оба шага могут объединить данные с экспертными знаниями за счет использования подходящих априорных распределений по G и Θ (см., Например, Castelo and Siebes, 2000; Mukherjee and Speed, 2008; Druzdzel and van der Gaag, 1995). Если мы предположим, что параметры в различных локальных распределениях независимы и что данные не содержат пропущенных значений (Heckerman et al., 1995), мы можем выполнять обучение параметрам независимо для каждого Xi

Finally, once both and have been learned, we can answer queries about our quantities of interest using the resulting BN as our model of the world. Common types are conditional probability queries, in which we compute the posterior probability of some variables given evidence on others; and most probable explanation queries, in which we identify the conﬁguration of values of some variables that has the highest posterior probability given the values of some other variables. The latter is especially suited to implement both prediction and imputation of missing data. These queries can be automated, for any given BN, using either exact or approximate inference algorithms that work directly on the BN without the need for any manual calculation; for an overview of such algorithms see Scutari and Denis (2014).

Наконец, когда оба G и Θ были изучены, мы можем ответить на вопросы о наших интересующих величинах, используя полученную BN в качестве нашей модели реальности. Распространенные типы - это запросы с условной вероятностью, в которых мы вычисляем апостериорную вероятность одних переменных с учетом данных о других; и запросы наиболее вероятного объяснения, в которых мы идентифицируем конфигурацию значений некоторых переменных, которая имеет наивысшую апостериорную вероятность при значениях некоторых других переменных. Последний особенно подходит для реализации как прогнозирования, так и включения недостающих данных. Эти запросы могут быть автоматизированы для любой заданной BN с использованием алгоритмов точного или приблизительного вывода, которые работают непосредственно с BN без необходимости в каких-либо ручных вычислениях; обзор таких алгоритмов см. в Scutari and Denis (2014).

BNs can model a wide variety of phenomena effectively. However, their applicability is not always apparent to practitioners in otherﬁelds due to the strong focus of the literature on the simple scenario in which data are static (as opposed to dynamic, that is, with a time dimension) and complete (as in, completely observed). Dynamic data are central to a number of cutting-edge applications and research in ﬁelds as different as genetics and robotics; and incomplete data are a fact of life in almost any real-world data analysis. BNs can handle both in a rigorous way, as we will see in the following.

BN могут эффективно моделировать самые разные явления. Однако их применимость не всегда очевидна для практиков в других областях из-за сильного внимания в литературе к простому сценарию, в котором данные являются статическими (в отличие от динамических, то есть с временным измерением) и полными (например, полностью наблюдаемый процесс). Динамические данные занимают центральное место в ряде передовых приложений и исследований в таких разных областях, как генетика и робототехника; а неполные данные - это реальность практически любого анализа реальных данных. Как мы увидим ниже, BN могут обрабатывать и то, и другое.

Dynamic BNs (DBNs ) combine classic (static) BNs and Markov processes to model dynamic data in which each individual is measured repeatedly over time, such as longitudinal or panel data. An approachable introduction is provided for instance in Murphy (2002). They have major applications in engineering (Pavlovic et al., 1999; Frigault et al., 2008), medicine (van der Heijden et al., 2014), genetics and systems biology (Perrin et al., 2003). The term “dynamic” in this context implies we are modelling a dynamic system, not necessarily that the network changes over time.

Динамические BN (DBN) объединяют классические (статические) BN и марковские процессы для моделирования динамических данных, в которых каждый индивидуальная величина измеряется многократно с течением времени, например, долготные или координатные данные. Доступное введение дается, например, в Murphy (2002). Они находят широкое применение в технике (Павлович и др., 1999; Фриго и др., 2008), в медицине (van der Heijden et al., 2014), в генетике и системной биологии (Perrin et al., 2003). Термин «динамический» в этом контексте подразумевает, что мы моделируем динамическую систему, не обязательно, чтобы сеть изменялась с течением времени.