第 28 卷 增刊
2002 年 12 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 28, Suppl.
Dec., 2002

机器智能与模式识别研究中的统计学习方法[1)]

王天树 郑南宁 袁泽剑

(西安交通大学人工智能与机器人研究所 西安 710049)

(E-mail: nnzheng@mail.xjtu.edu.cn)

摘 要 简要介绍了学习算法的发展状况;讨论了机器智能与模式识别研究中的统计学习方法和图模型的一般理论;重点叙述了图模型的统计推断过程和学习算法以及应用统计学习方法解决问题的一般步骤;最后给出了用于时间序列分析的动态贝叶斯网络的实例.

关键词 机器智能,模式识别,图模型,统计学习

中图分类号 TP18

STATISTICAL LEARNING IN MACHINE INTELLIGENCE AND PATTERN RECOGNITION

WANG Tian-Shu ZHENG Nan-Ning YUAN Ze-Jian

(*Institute of the Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an* 710049)

(E-mail: nnzheng@mail.xjtu.edu.cn)

Abstract In this paper, the general theory of statistical learning methods and graphical model in machine intelligence and pattern recognition are discussed. We specially depict the inference process and learning algorithm of graphical model, and the general steps of application of statistical learning methods to solve problems. In the last, we give an example of dynamical Bayesian networks for time series data analysis.

Key words Machine intelligence, pattern recognition, graphical model, statistical learning

1 引言

自从 20 世纪 40 年代计算机出现以来,随着科学技术的进步,计算机的计算能力得到了极大的提高,并且在人类的日常生活中扮演着越来越重要的角色.但迄今为止,计算机对自然信息处理的能力仍然十分有限.如何使计算机具有与人同样的信息处理能力是机器智能研究的主要目标[1].

长期以来,统计学习方法一直是机器智能领域的研究重点之一.特别是近 10 年,人工智能领域的学者正试图利用统计学习框架来统一已有的在机器智能领域所提出的若干模型和

1) 国家创新研究群体科学基金(60024301)、国家自然科学基金(60175006)资助

收稿日期 2002-04-20 收修改稿日期 2002-09-03

方法[2,3]. 1962 年 Rosenblatt 提出了称作感知机(perceptron)的神经网络模型[4],感知机是一种能够进行自学习,并能够在自学习的基础上进行归纳判断的计算装置. Rosenblatt 还给出了数学上的分析和数字计算机的模拟. 这标志着人们对学习问题进行数学研究的开始. 当时感知机的出现以及 Rosenblatt 的成功的工作促使众多的研究者投入到人工神经网络的研究中. Minsky 和 Papert 在 20 世纪 60 年代后期对感知机的功能作了更进一步的数学分析. 两人合作于 1969 年发表了一本颇有影响的《Perceptrons》专著,书中指出感知机不能实现基本的异-或(XOR)功能,因此不是一个适当的模型. Minsky 是人工智能学科的创始人之一,由于他在学术界的地位和影响,使得许多学者接受了他和 Papert 的这一悲观结论,从而造成当时的人工神经网络研究一时出现低潮,许多研究处于中止. 随后的 60～70 年代,尽管 Minsky 和 Papert 对于人工神经网络研究持否定态度,但人工神经网络研究仍然在小范围内继续,学习算法得到了很大发展. Grossberg 研究了认知、记忆和视觉的神经激活机理,他应用人工神经网络来进行人脑思维的模型化. Grossberg 的人工神经网络模型特性的数学分析对认知和记忆做了适当的解释. 1971 年 Amari 提出了布尔人工神经网络理论,即人工神经网络中仅含有布尔值;Anderson 提出了完全分布的线性联想记忆;Kohonen 则提出了自组织联想记忆,他研究了人脑究竟是怎样组织其记忆的信息. 在 60～70 年代,人们还提出了一系列的学习机,如 Widrow 构造的 Madaline 自适应学习机[5]、Steinbuch 提出的学习矩阵(learning matrix)[6]等. 为了解决一些应用领域的实际问题,人们还提出了隐马尔科夫模型(Hidden Markov Model,HMM)[7]等具体算法. 60 年代,统计学习的基本理论也得到了很大发展. 例如,Vapnik[8]提出了 VC 维数、VC 熵等统计学习中的基本概念;Tikhonov[9]等提出了求解不适定问题的正则化算法;Parzen[10]等人研究了密度估计的非参数化方法;Kolmogorov[11]等人给出了算法复杂度的基本思想. 这些工作在统计计算领域中产生了深远的影响,并极大地推动了模式识别与机器智能中学习方法的研究.

在 20 世纪 70 年代后期,由于人工神经网络在认知模型中的应用,使人工神经网络技术在认知心理学领域中开始具有显著的地位. 推动这一研究的有两位著名的认知心理学家,他们是 UCSD 的 Rumelhart 和 Carnegie-Mellon 大学的 McClelland. 他们两人工作的重要意义在于提出了并行分布处理(Parallel Distributed Procession,PDP)模型. 许多人工神经网络学习算法都采用了 PDP 模型,如竞争学习算法,波尔兹曼(Boltzmann)机和误差传播(error propagation)算法.

从 20 世纪 80 年代开始到 90 年代初期,产生了一批重要的神经网络模型. 1982 年 Hopfield 在一份向美国国家科学院提交的研究报告中提出了“计算能量函数”的概念[12],给出了神经网络的稳定性判据. Hopfield 网络模型还可用于联想记忆和各种优化计算,开拓了神经网络用于计算的途径. 同年 Kohonen[13]发表了讨论自组织映射(Self Organizing Map,SOM)的论文;在 1984 年 Kirkpatric[14]针对组合优化问题提出了模拟退火算法,随后模拟退火的基本思想被用于 Boltzmann 机[15]. 特别是 1986 年 Rumehart 等人[16]提出的用于多层感知器学习的反向传播(BP)算法引发了神经网络的研究热潮. Kosko 提出了模糊认知映射系统[17],模糊认知映射可以处理不清楚、非确定性(模糊数据)、矛盾的和错误的数据,非常适合涉及到个体交互的知识基础上的复杂系统,这种方法已经应用在雷达图像处理中. 将模糊计算与神经网络模型相结合是一种很好的软计算方法[18,19]. 人工神经网络在控制、模式识别与机器智能等领域得到了成功的应用. 但是,不同的神经网络模型基于不同的数学理论,很难建立一套统一的理论体系结构.

从20世纪90年代到现在，人们开始发展与统计学习相关的一些方法[20,21]. 由于统计学习的基本理论框架已经在60～70年代建立，近年来的研究重点在于提出一些解决问题的新算法和新模型[22～25]，而所有的算法和模型均以统计学习理论作为统一理论基础. 其中比较重要的研究工作如：Comon[26]在1994年首先提出的独立元分析(Independent Component Analysis，ICA)算法；1995年Vapnik[27]等人提出的支撑向量机(Support Vector machine，SVM). 随后支撑向量机又被扩展到核学习(kernel learning). 目前已经有相当多的基于核学习的研究工作[28]，这些工作提出的模型均为非线性模型. 虽然非线性模型具有比线性模型更强的表述能力，但是由于研究非线性系统的数学理论还不成熟，完善核学习的理论框架将是一个长期艰苦的过程. 与此同时，利用贝叶斯(Beyesian)学习的统计建模方法也得到了很大的发展. 特别是EM[29]，MCMC[30]和RJMCMC[31]算法的引入，使得Beyesian机——Beyesian分析和Beyesian统计计算的结合[32]成为处理数据和知识的一种强有力的工具.

2　统计学习模型

从样本进行学习的一般模型包括三个部分：样本产生器、训练器和学习机器[20,21]. 学习的问题就是从给定的函数集中选择出能够最好地逼近训练器响应的函数. 尽管学习问题的形式描述极其广泛，但主要集中在模式分类(classification)、回归估计(regression)和密度估计(density estimation)这三种主要问题上. 其中密度估计问题是统计学习中最基本的问题. 如果解决了密度估计问题，就可以得到联合概率密度. 而模式分类和回归估计计算的是条件概率密度，条件概率密度可以通过边界化联合概率密度得到. 统计模型的形式与学习问题与所在的应用领域密切相关.

2.1　基于样本统计特征的统计模型

通常观测数据的维数很高，直接对其建立统计模型十分困难，因此需要在样本数据上提取不同的特征. 对于一个数据集合，可以依据最大熵原理构造与观测数据统计特性一致的吉布斯(Gibbs)分布[33]. 对此分布采样能得到的新的样本集合. 而从新样本集合提取的特征的统计特性与训练数据完全一致. 这种模型是最一般的统计模型，具有构造型的特点，它等价于图模型中的马尔科夫网络(Markov network).

2.2　应用贝叶斯网络(Beyesian network)和能量函数近似估计(energy function approximation)的统计模型

直接利用样本统计分布采样的统计模型的计算复杂度很高，主要是在吉布斯模型中的配分函数(partition function)的计算复杂度与模型中的节点个数呈指数关系. 这样直接计算配分函数是一个NP hard问题. 为了解决这一问题，可以采取两种方案. 其中之一是使用有向图结构的贝叶斯网络来近似原始的统计模型. 由于对近似模型可以直接计算，因此能间接地得到近似结果. 另外一种方法是仍然使用无向图结构，但对每个节点引入信用(belief)的概念. 信用利用模型中某一个局部包含的节点计算，从而避免计算全局的配分函数的计算. 利用信用的传播更新过程就可以逼近原来的统计模型. 这种方法来源于统计物理方法，信用更新过程与统计物理中的能量函数优化过程等价. 属于这一类的方法有均值场(mean field)，Bethe和Kikutch能量函数[34]等. 利用这些近似过程，原来在有向图上提出的信用传播算法(belief propagation)可以在无向图上得到很好的近似结果. 但是信用传播方法目前只适用于包含离散节点的马尔科夫网络.

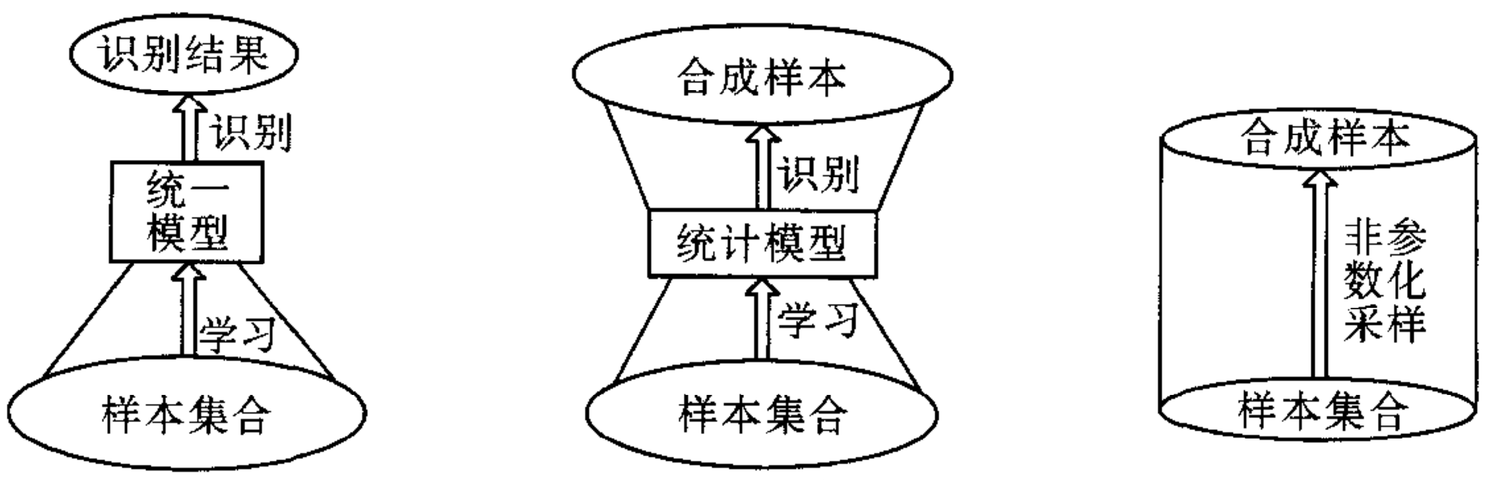
2.3 产生式模型(generative model)

在利用基于样本统计特征的统计模型进行统计建模时,有时无法给出产生样本数据的内在机制.因此模型类似于一个黑箱,需要对模型包含的隐含变量和其统计相关性作适当的假设,并认为观测数据来源于产生这些隐变量的随机过程.这种建模方法称为产生式模型.产生式模型的学习过程就是估计隐变量的分布和描述其相互关系的参数辨识的过程.通常产生式模型具有清晰的分层结构,而学习得到的模型很容易满足模型解释要求.

2.4 基于条件概率密度的统计模型

直接利用样本统计特征的统计模型的学习过程往往缺乏效率,因为密度估计是学习问题中最困难的问题.对于特定的学习任务,如模式分类和回归分析,可以避开密度估计问题,直接针对问题设计特定的学习算法.这时,通常只需要学习条件概率密度,从而简化了计算过程以及准确估计模型参数所需要的样本数据量.直接描述条件概率密度的模型又称为判别式模型(discriminative model).这类模型的典型代表是支撑向量机(SVM)[21].如果以识别为学习的目的,学习得到的模型需要尽量从样本数据中抽取共有的特征,以得到正确的分类边界.这样的模型通常属于判别式模型,它并不包含单一样本的具体特性.由于这类模型一般并不包含足够的描述样本真实分布的信息,针对识别设计的模型和方法不再适用计算机视觉中的动态场景合成的学习.

产生式模型直接估计样本服从的统计分布,通过对学习得到的分布采样就可以直接得到合成的样本.而密度估计问题是学习中最难解决的问题,它是典型的病态(ill-posed)问题,而且很多具体应用中没有足够的样本来估计一个复杂模型的参数.Efros[35]等人提出的非参数化纹理合成方法,使人们注意到非参数化模型在合成领域的应用.从样本到样本的直接采样过程避免了复杂的参数学习过程,同时能很好地保持样本的统计特性.非参数化采样方法是目前合成纹理的最佳方法[36,37].非参数化采样方法对训练样本集的依赖性很强,样本集必须包含能描述数据统计特性的足够样本.用于识别和合成的不同模型的比较如图 1 所示.



(a) 用于识别的统计模型 (b) 用于合成的统计模型 (c) 非参数化采样

图 1 不同统计结果的比较

用统计方法合成样本时,统计模型的熵越小,采样得到的样本就越"象"原始样本集.然而另一方面,统计模型需要包含足够的随机性来产生丰富多变的样本.实际的统计模型需要在二者之间折衷.在利用统计学习解决模式识别问题时,需要紧密结合相关应用领域的知识来选择合适的模型结构[38,39].

3 统一的概率图模型(图模型)

图模型(graphical model)理论是图论(graph theory)和统计理论相结合的产物.1988 年

Pearl 在他的开创性著作[40]中系统地阐述了利用概率网络(pro-babilistic networks)在人工智能和专家系统中描述不确定性的方法.随后的十多年间,概率网络也就是贝叶斯网络在理论和应用方面都得到了巨大的发展[41,42].贝叶斯网络可以用有向图表示.但在贝叶斯网络的基础上发展起来的学习算法,也适用于结构为无向图的马尔科夫网络(Markov network). 对这些以图形表示的概率模型,可以统称为概率图模型.因此,在某种程度上,图模型理论统一了在统计、系统工程、信息论、模式识别和统计物理中得到广泛应用的统计模型,如高斯混合模型、主成分分析(PCA)、独立元分析(ICA)、波尔兹曼机(Boltzmann machines)、因子分析(FA)、混合因子分析、隐马尔科夫模型(HMM)、因子化的隐马尔科夫模型(factorial HMM)、混合的隐马尔科夫模型、线性动态系统、混合的线性动态系统、可切换的状态空间模型(switching state-space models)、非线性动态系统等.

图模型的基本思想是把一个复杂系统按照不同的层次结构分割成一些基本的组成部分.这些组成部分的依赖关系用图形表示.而使用统计理论可以把这些部分组合成为一个系统,同时保证各个部分在整个系统中协调一致.通过分解过程就可以把复杂问题转化为一组简单的问题.因此,使用图模型可以很好地解决模式识别和机器智能研究中的不确定性和复杂性的问题.同时,基于图模型的统一的表述方法有助于设计通用、高效的推理和学习算法.

在图模型中,每一个节点代表一个随机变量.随机变量之间的关系用图模型中节点之间的有无联线加以表示.两节点间没有联线表示这两个节点所代表的随机变量之间没有直接的关系,或者说这两个随机变量是条件独立的.根据连线是否具有方向,可以把图模型分为基本的两类:一类为有向图模型;另一类为无向图模型.既有有向连线又有无向连线的图模型为链图模型.在无向图模型中,随机变量之间具有对称关系.无向图模型中统计计算往往是建立在阀值集(cliques)的基础上,因此也称之为马尔科夫网络(Markov network).马尔科夫网络的典型实例是在初级视觉问题得到了广泛应用的马尔科夫随机场(Markov Random Field, MRF)[33].相反,在有向图模型中,节点之间的连线是有方向的,这种方向性刻画了随机变量之间的条件关系.由于在有向图模型中推理是建立在贝叶斯律基础上的,因此有向图模型也称为贝叶斯网络.贝叶斯网络相对马尔科夫网络作了更强的条件独立性假设,在一定的条件下也具有相应的高效统计推断和学习算法.值得注意的是,贝叶斯网络的名称来源于其统计推断过程中使用的贝叶斯规则,但并不意味着贝叶斯网络的学习过程一定是贝叶斯学习.

一个简单的贝叶斯网络如图 2 所示.图中包含 6 个节点.如果两个节点之间存在连线,则这两个节点分别被称为父节点和子节点,箭头指向的为子节点.这里记任意一个节点 X_i 的所有的父节点集合为 X_{π_i}.如在图 2 中,节点 X_6 的父节点集合 $X_{\pi_6}=\{X_2,X_5\}$.

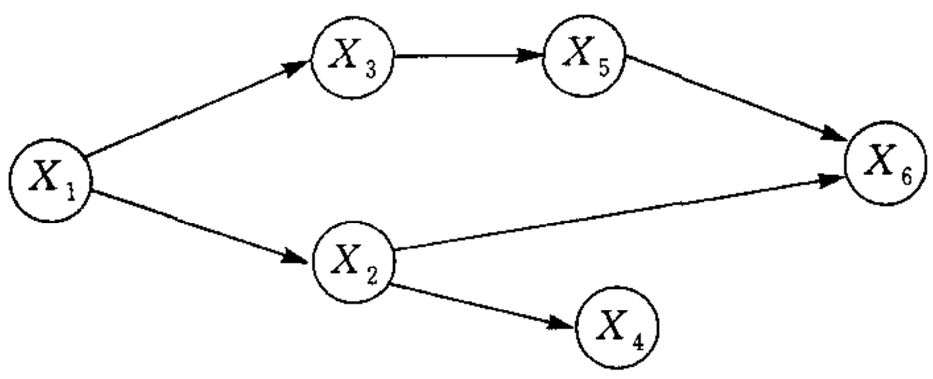


图 2　简单的贝叶斯网络

在图 2 贝叶斯网络中,6 个随机变量之间的关系蕴涵于其联合概率密度中.利用随机变量之间的条件独立关系,一个贝叶斯网络的行为完全可以用随机变量的局部条件概率来刻画.一个复杂的联合概率密度也就简化为局部条件概率的乘积,即

$$p(x_1,x_2,x_3,x_4,x_5,x_6)=$$
$$p(x_1)\cdot p(x_2|x_1)\cdot p(x_3|x_1)\cdot p(x_4|x_2)\cdot p(x_5|x_3)\cdot p(x_6|x_2,x_5) \quad (1)$$

更一般地,对任何 n 个节点贝叶斯网络,其包含的所有节点的联合概率密度也可以用其局部

条件概率来表示,即

$$p(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}p_i(x_i|x_{\pi_i}) \tag{2}$$

其中 $p_i(x_i|x_{\pi_i})$表示第 i 的节点的条件概率密度.

利用随机变量的条件分布独立关系来简化联合概率分布,可以大大减少模型包含的参数.确定一个贝叶斯网络不仅需要指定网络的图形结构,完整的模型还需要为节点的局部条件概率指定具体统计结构形式,即需要确定条件概率密度 $p_i(x_i|x_{\pi_i})$.选择合适的条件概率密度形式与发现图模型中的子结构或拓扑结构是图模型理论研究中的核心问题之一.

图模型是对随机变量之间关系的一种直观的描述.建立概率图模型的目的是能够借助于图理论和统计理论来分析随机变量之间的内在的联系,也就是说希望能通过观测数据或部分观测数据来发现隐含在观测数据后的规律性.对随机变量之间的关系分析主要体现在对隐随机变量的推理和对图模型结构以及节点与节点之间的映射关系的学习上.图模型是对经典的统计模型的统一描述,因此传统的统计推断算法[30]、模型的选择与模型参数学习算法[29]也适用于所有图模型的学习与推理问题.同时,建立在图模型框架下的学习算法也适用于经典的统计模型中的学习问题.

4 统计推理与学习方法

图模型是一组随机变量的联合分布的直观的描述.建立在图模型基础上的学习任务就是根据观测数据来得到其联合分布[3].图模型的统计推断过程也是在联合分布的基础上展开的.在图模型中所有的节点可分割成互不相交的子集 X_E,X_F,X_H. 其中 X_E为观测数据的节点或证据节点,X_F为统计推断中需要计算后验概率的节点(其后验概率可以用于解决决策、识别等问题),X_H为图模型中包含的隐节点.统计推断过程就是计算后验概率分布 $p(X_F|X_E)$的过程.图模型的学习是获取图模型拓扑结构与节点局部条件概率参数(模型参数)的过程.一般情况下,模型结构的学习要比模型参数的学习困难得多.

4.1 图模型统计推理

图模型统计推断过程是利用观测节点的分布计算其他节点的条件概率的过程.即计算 $p(X_F|X_E)$.按照条件概率公式可以得到

$$P(X_F|X_E)=\frac{\int p(X_E,X_F,X_H)\mathrm{d}X_H}{\iint p(X_E,X_F,X_H)\mathrm{d}X_H\mathrm{d}X_F} \tag{3}$$

直接利用这一公式计算条件概率涉及到高维的积分运算问题,对复杂模型无法直接应用.

研究统计推断过程的中心问题就是提出高效的计算方法.Pearl[43]在 1983 年提出了利用消息传递机制实现统计推断过程的方法.他的方法只适用于单连接贝叶斯网络.该算法在一般贝叶斯网络上的推广就是 Junction Tree(或者 Join Tree, JT)算法.Lauritzen[44], Shafer[45]等人分别研究了 Junction Tree 算法的两种不同实现.之后,JT 算法得到了不断改进[22],并已成为图模型统计推断过程中的标准算法.JT 算法可以被看作是一个消息传递的过程.贝叶斯网络中的每个节点首先按照一定次序向其它节点传递描述自己概率分布的信息.之后,所有的节点利用收集到的信息更新自己的后验分布.JT 算法包含以下过程.

1) 规范化(moralization)

马尔科夫网络的联合概率分布密度也可以被分解为簇(两两相连的节点集合)C_i的势能函数 ψ_{C_i}乘积. 由此可以把贝叶斯网络的局部条件概率分布转化为对应的马尔科夫网络的势能函数,同时把有向图变为无向图,这样就可以把图模型的学习统一起来.

2)构造连接树(junction tree)

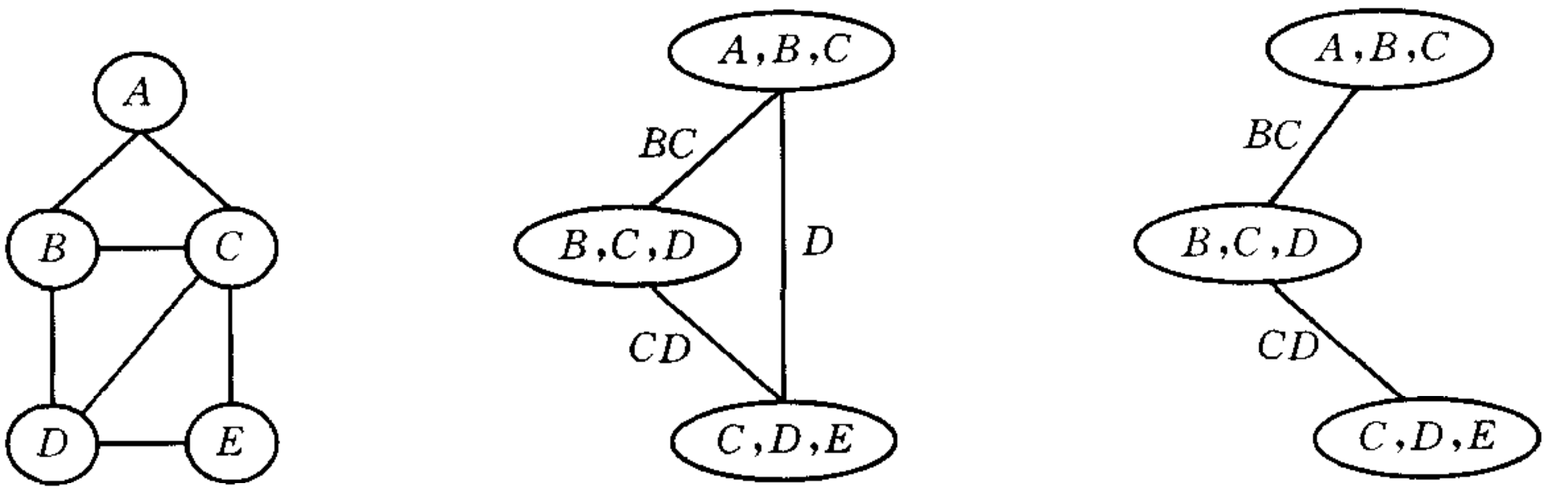
统计推断过程需要计算当观测节点取值 $X_E=e$ 时的节点集合 X_F的条件概率密度

$$p(X_F|X_E=e)=\prod_C \psi_C(X_F|X_E=e)=\frac{\prod_C \psi_C(X_F,X_E=e)}{\sum_{X_F}\prod_C \psi_C(X_F,X_E=e)} \tag{4}$$

计算(4)式的条件概率需要对所有节点求和. 为了提高计算效率,需要找到将其分解的策略. 可以证明条件概率分布可以分解为边缘分布之积

$$p(X_F|X_E=e)=\frac{\prod_C \psi_C}{\prod_S \phi_S}=\frac{\prod_C p(X_C,X_E=e)}{\prod_S p(X_S,X_E=e)} \tag{5}$$

其中 ψ_C为一个簇 C 包含的节点 X_C的边缘分布,ϕ_S为簇的分隔(seperator,两个簇之间的公共部分)包含的节点 X_S的边缘分布,在初始化时取常数值 1. 由于边缘分布只在一个簇包含的节点上计算,效率得到大大提高. JT 算法就是计算 ϕ_S和 ψ_C过程. 一个构造 Junction Tree 的示例如图 3 所示.



(a) 无向图　　(b) 簇和分离　　(c) Junction Tree

图 3　构造最大张成树过程示例

3)消息传递与更新 ψ_C过程

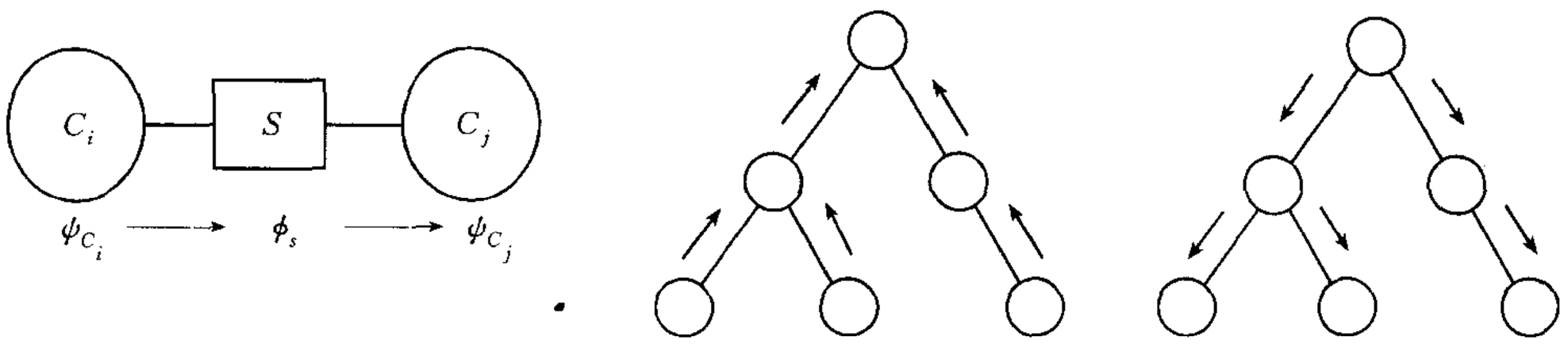
在模型规正过程中已经给 ψ_C指定了初值,但是 ψ_C只包含局部的信息. 对两个相邻的簇,分别用各自的 ψ_C计算其公共部分的边缘分布,取得的结果可能会不一致. 为了使模型协调一致,需要利用消息传递过程更新 ψ_C. 对于已经表示成树结构的图模型,消息传递过程包含以下两步:

a) 证据收集(collect evidence),从叶节点开始,每个子节点收集到自己所有的子节点传来的消息,利用此消息更新自己的分布,之后向它的父节点传递消息,直到根节点为止;

b) 证据派发(distribute evidence),从根节点开始,每个父节点收集到自己所有的父节点传来的消息后,更新自己的分布,紧接着向它的所有子节点传递消息,直到叶节点为止.

消息传递过程的示例如图 4 所示. 经过消息传递过程,可以得到图形中所有节点的边缘分布,这样就完成了统计推断的计算. 经典的前向后向(foward-backward)[46]统计推断算法可以被看做是 JT 算法的一个特例.

在复杂模型中应用 Junction Tree 算法时,可能会遇到包含大量节点的簇. 这时仍然无法直接计算节点的后验概率分布. 在实际应用中,对这类复杂的情况可使用近似统计推断算



(a) 消息传递 (b) 证据收集 (c) 证据派发

图 4 图模型中的消息传递过程

法. 常用的近似算法包括 MCMC 采样[23]、变分逼近 (variational approximation)[24]等.

4.2 图模型的学习

图模型的学习是指学习模型结构(或拓扑)、参数,或者对模型结构与参数同时进行学习. 在给定观测数据的情况下,可以利用相应的学习算法来获取模型的结构及相应的参数. 在一般情况下,模型结构的学习比参数的学习要困难得多. 如果图模型中含有隐节点时,学习问题就会变得更加困难. 按照学习过程的难易程度,可把图模型的学习分为四种情况[47]: 1)图模型结构已知,全部节点都是可观测节点;2)图模型结构已知,包含隐节点;3)图模型结构未知,全部节点都是可观测节点;4)图模型结构未知,包含隐节点. 在学习算法中,参数估计的极大似然方法与能够进行模型结构与参数估计的贝叶斯方法是最基本的. 极大似然估计的计算与贝叶斯计算也是统计领域中两类主要计算问题[48]. 在实际应用中的许多学习算法都是在此基础上派生出来的[49~51].

4.2.1 极大似然学习

在图模型结构 M 已知,从观测节点得到的样本集为 Y 的情况下,对模型参数 θ 进行极大似然估计可描述为如下优化问题

$$\hat{\theta} = \arg\max_{\theta}\{P(Y;\theta)\} \tag{6}$$

当图模型中包含隐节点时,模型参数的学习会变得复杂. 解决带有隐节点的极大似然参数估计的经典方法是 EM 算法[29], 以及在 EM 算法上的扩展[49,51]. EM 算法把一个复杂的极大化问题转变为一系列简单的极大化问题. EM 算法的每一个迭代周期包含两步:在 E 步(Expectation),固定当前模型参数,使用统计推断算法,结合观测数据和计算隐节点的分布;在 M 步(Maximization),固定 E 步得到的隐节点分布,利用最大似然准则更新模型的参数. 在图模型中,设隐变量集合为 X,图模型的联合概率为 $P(X,Y;\theta)$. 给定观测节点上的观测样本,相应的似然函数的对数为

$$L(\theta) = \log\int P(X,Y;\theta)\mathrm{d}X \tag{7}$$

由于在 $L(\theta)$中,模型的参数之间存在很强的耦合关系,直接最大化 $L(\theta)$是十分困难的. 通常方法是利用 Jensen 不等式将 $L(\theta)$的直接处理转化为对其界的处理,即

$$L(\theta) = \log\int Q(X)\frac{P(X,Y;\theta)}{Q(X)}\mathrm{d}X \geqslant$$

$$\int Q(X)\log P(X,Y;\theta)\mathrm{d}X + H(Q) = F(Q,\theta) \tag{8}$$

其中 $Q(X)$为隐变量的一个分布,$H(Q)$为分布 $Q(X)$的熵,$F(Q,\theta)$为 $L(\theta)$的一个下界.

为了减少逼近误差,隐变量的分布 $Q(X)$应该使下界尽可能紧,也就是说选择一个隐变

量的分布使得下界极大化. 在理想情况下，隐变量的分布为 $Q(X)=P(X|Y;\theta)$，即使得(8)式中的等式成立，但在复杂的图模型中这种情况是很难满足的. 给定一个初始的模型参数 θ^0，EM 算法的过程是交替如下两步

$$\text{E Step: } Q^{k+1}=\arg\max_{Q}F(Q,\theta^k),$$

$$\text{M Step: } \theta^{k+1}=\arg\max_{\theta}F(Q^{k+1},\theta),$$

来极大化下界 $F(Q,\theta)$的过程. 在极大化下界的过程中，E 步总是寻求一个隐变量的分布在当前模型参数值的情况下尽可能逼近其真实分布；M 步是更新当前的模型参数使似然函数的值增加. 在多数情况下，极大似然估计是一个非凸优化问题，EM 算法得到的是一个局部最优解，并且敏感于初始的模型参数. 为了改善 EM 算法的性能，文献[52,53]提出了一些改进算法.

4.2.2　贝叶斯学习

当模型的结构不确定时，极大似然学习由于没有考虑模型的复杂性，从信息论的角度来说，即没有考虑对模型参数进行编码所花费的代价，从而会引起过拟合问题. 因此，极大似然学习方法不能够进行模型选择和确定模型结构. 贝叶斯方法从理论上能解决参数过拟合问题，并能够进行模型学习. 在贝叶斯方法中，模型参数是随机变量. 模型学习是指给定观测数据去寻找合适的模型或模型结构. 在模型 M 条件下，观测节点集 Y 的边缘概率为

$$P(Y|M)=\int P(Y|\theta,M)P(\theta|M)\mathrm{d}\theta \tag{9}$$

其中 $P(\theta|M)$为模型 M 参数的先验概率. 贝叶斯学习方法最初是从简单的知识开始，这些简单的知识是指模型的分布 $P(M)$以及与模型相对应的参数先验分布 $P(\theta|M)$，然后用观测数据来更新模型和参数的先验分布，这些更新后的先验分布用模型和参数的后验分布来表示. 利用贝叶斯公式有模型的后验概率

$$P(M|Y)=\frac{P(Y|M)P(M)}{P(Y)} \tag{10}$$

参数的后验概率分布为

$$P(\theta|Y,M)=\frac{P(Y|\theta,M)P(M|\theta)}{P(Y|M)} \tag{11}$$

在模型和参数的后验概率的基础上，选取不同的损失函数就可以对模型和参数作出决策，如最大后验估计(MAP)、后验均值估计(PM)等.

尽管贝叶斯方法在理论上可以避免过学习问题和进行模型选择与平均，但在实际应用中会涉及到难以处理的高维积分问题：

1)如在(9)式中沿参数矢量的高维积分；

2)在计算关于观测数据的概率时涉及到模型的平均

$$P(Y)=\sum_{M}P(Y|M)P(M) \tag{12}$$

鉴于精确计算这些高维积分的困难，在实际应用中往往使用一些逼近算法. 这些算法大致分为两类：其一是确定性逼近算法；其二是随机模拟的算法. 在确定性逼近算法中，得到广泛应用的是拉普拉斯逼近，以及在此基础上定义的一些模型选取准则包括[50]最小描述长度(Minimal Descriptive Length，MDL)、贝叶斯信息准则(Bayesian Information Criterion，BIC)、最小信息长度(Minimal Message Length，MMI)等. 在随机模拟逼近算法中最有影响的是 MCMC 方法，以及可在模型空间中进行跳跃的 RJMCMC 方法[31]. 模拟方法尽管具有

很高的逼近精度,但此算法在时间和空间所花费的代价极高,对于 MCMC 方法来说还会涉及到马尔科夫链的收敛性问题. 最近,变分方法在模型选择和参数估计中也受到关注,它是一种在逼近精度和算法复杂度之间较为折中的逼近方法[54].

5 应用统计学习方法解决问题的步骤

利用统计学习方法获得描述数据的统计模型只是解决具体问题中的一步. 要得到问题的正确求解,需要正确选择统计学习方法中具体过程的每一个步骤. 一个完整的利用统计方法解决问题的过程包含以下步骤[20].

1)问题的定义. 大多数统计模型和方法是针对特定领域的应用问题提出的. 为了给出有意义的问题定义,首先要考虑特定领域的知识和经验. 不幸的是,近期的一些研究工作往往只注重于讨论学习算法,而忽略了如何准确清楚地定义问题.

2)收集实验数据. 实验数据的来源可能是受控的也可能是不受控的. 所谓受控是指可以"设计"实验,即可以通过设定实验条件,选择适当的实验数据;而不受控时,就只能处于观测者的地位. 统计学习过程通常面对的是数据不受控的情况,即实验数据样本是通过对某个分布进行随机采样获得的. 明确数据产生的机制也很重要,因为通过不同采样方法得到的实验数据,可能导致不同的学习结果.

3)数据预处理. 数据预处理是一个非常重要的环节,它影响到整个学习过程的成败. 对观测样本的预处理一般(至少)包含两类内容:检测和消除非正常的离群样本(outlier),以及对样本做适当变换. 离群样本(outlier)是指与大多数观测样本统计特性不一致的数据. 这些离群样本可能来自不正确的观测过程或者观测样本在编码传输过程中出现了错误. 这样的数据严重影响学习算法的性能,必须加以消除. 可能的解决办法有两类:检测并去除非正常数据,以及使用鲁棒统计(robust statistics)方法[55].

对数据进行适当变换在实践过程中也是非常重要的环节,它包括特征选择(feature selction)、降维(dimension reduction)、变量归一化(normalization)等. 在一般情况下数据变换的目的是解决样本数据维数过高及不同变量之间单位不一致的问题. 数据维数过高时,可以使用特征选择或其它一些降维方法. 变量数值范围相差过大时,可以进行适当的归一化. 注意某些学习算法性能会受到归一化过程的影响.

4)选择或设计模型. 对同一个问题,往往可以使用不同的统计模型来描述. 因此需要充分利用该领域已有的专家经验和知识,选择合适的统计模型. 除此之外,还需要确定模型的具体形式,这不仅需要选择待学习系统的合适输入和输出变量,还需要(如果可能)设定系统内部的随机变量以及它们相互之间依赖关系的一般形式.

5)学习模型参数. 有了第 4)步定义的模型结构和形式,就可以利用统计学习方法来估计模型的参数. 注意统计学习过程需要保证通过学习得到的模型对未知的数据有足够的适应能力[21],即模型要有足够的泛化能力. 在学习过程中需要避免模型过适定(over-fitting). 为了实现这一目的,可以利用贝叶斯学习办法,应用先验知识作为补偿项约束模型[2].

6)解释模型、验证模型. 通过学习得到的模型的一般用途是对未知数据做预测. 除此之外,往往还需要对模型的结构及参数做出解释. 因为模型的结构和参数体现了通过学习发现的产生样本数据机制,也可以说是通过学习获得的知识. 预测和解释这两者之间是矛盾的. 容易解释的模型结构相对比较简单,而能做出精确预测的模型则通常相当复杂. 传统的统计

学习方法偏向于使用结构较为简单的模型. 随着新的统计学习算法的发展，越来越多的复杂模型被提出来，以便得到对被学习系统更好的逼近[20].

以上列出的这些步骤对解决实际问题具有同等的重要性. 任何一步出现的问题都可能影响整个学习任务的成败. 统计学习方法并不是一种全能的方法，在解决实践问题过程中，还需要紧密结合特定问题所包含的领域知识. 利用这些领域知识可以在学习开始时对模型结构作适当假设，也可以在学习过程中作为评判模型优劣的标准.

6　动态贝叶斯网络实例

动态贝叶斯网络(Dynamic Bayesian NetWorks，DBNs)[56]是单连接(节点只有一个父节点)贝叶斯网络的一个特例. DBNs 是针对动态序列建模问题提出的. 一个典型的 DBNs 如图 5 所示. 从图中可以看出，DBNs 服从马尔科夫特性：t 时刻系统的所有变量只与$(t-1)$时刻系统的状态变量相关. 这是典型的马尔科夫链(Markov chain)[46]形式. 实际上，马尔科夫链可以作为 DBNs 的一个特例.

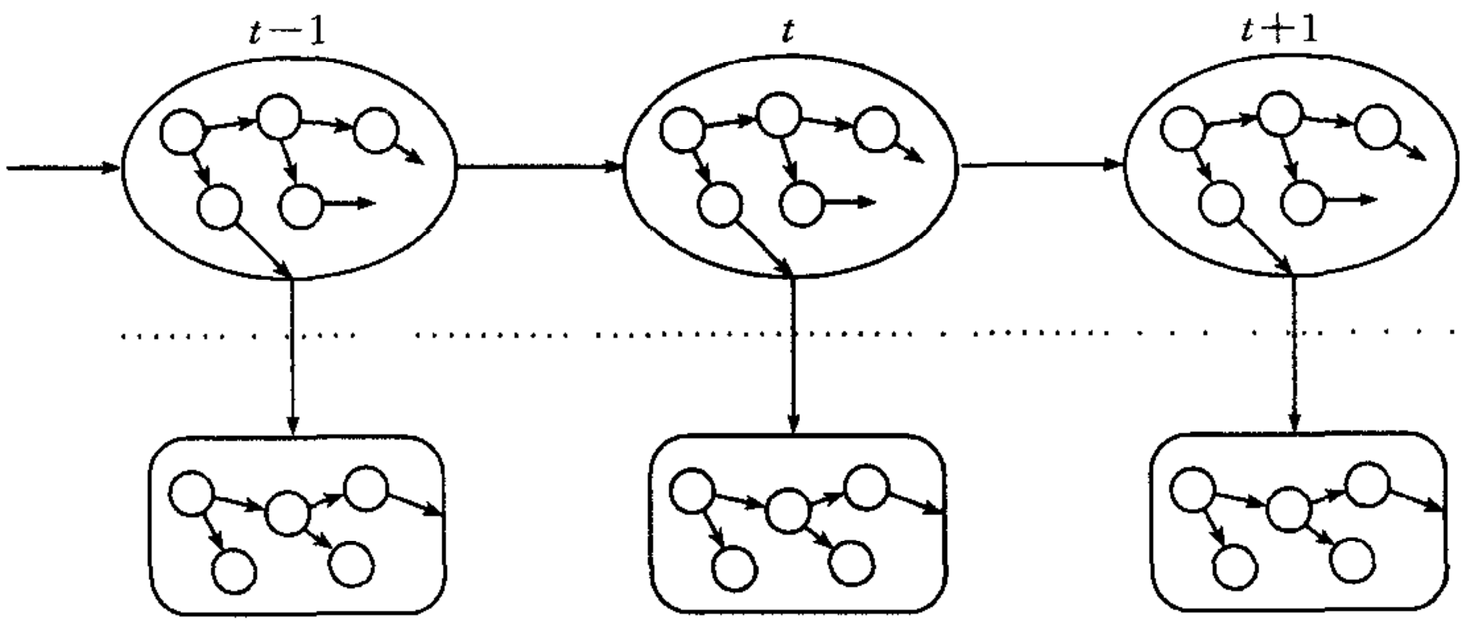
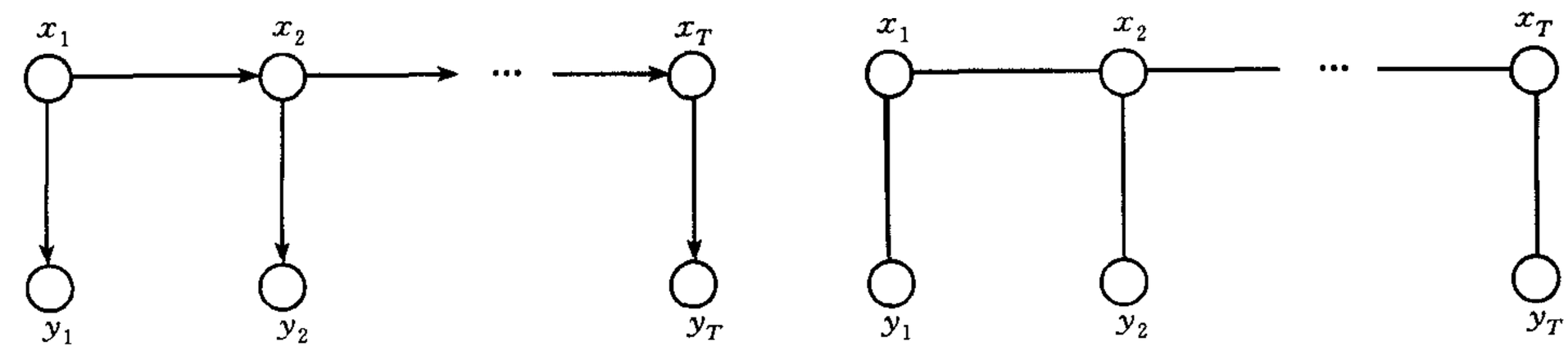


图 5　DBNs 的图模型表示

为了更清楚的表示动态贝叶斯网络，不失一般性，假设每个时刻 t 模型只包含一个随机状态变量 X_t和一个观测变量 Y_t. 这时图模型的拓朴结构如图 6 所示. 状态序列 $X=\{X_1,X_2,\cdots,X_T\}$和观测序列 $Y=\{Y_1,Y_2,\cdots,Y_T\}$的联合分布可以写成

$$P(X,Y)=\prod_{t=2}^{T}P(X_t|X_{t-1})\cdot\prod_{t=1}^{T}P(Y_t|X_t)P(X_t) \tag{12}$$



(a) 简化的 DBNs　　(b) 对应的无向图

图 6　简化为 DBNs 示例及对应的无向图

完全确定 DBNs 需要知道三个概率分布：状态转移条件分布$P(X_t|X_{t-1})$、观测条件分布$P(Y_t|X_t)$和初始状态分布 $P(X_1)$. 这里所有的条件分布可以是时变的或定常的，可以取参

数化形式 $P(X_t|X_{t-1};\theta)$，也可以使用非参数化(概率表格或统计直方图)表示. 根据状态变量是否连续，可以把 DBNs 分为离散 DBNs 和状态空间模型(state-space model). 其中离散 DBNs 的典型代表是隐马尔科夫模型(Hidden Markov Models，HMM)，状态空间模型的典型代表是线性动态系统(Linear Dynamic System，LDS)[55]. 与一般的图模型一样，在 DBNs 的学习过程中，首先需要确定统计推断算法. 有了统计推断算法就可以在 EM 框架下学习模型参数.

7 结束语

长期以来，科学和工程研究都建立在描述物理、生物、社会等系统的基本原理的基础上，在不同的应用领域利用基本的数学模型来解决具体的问题. 但是，在很多应用领域中，描述系统内在规律的基本原理是未知的，有时即使知道其基本原理，由于系统过于复杂也难以用数学模型精确描述. 其次，一个复杂系统的行为本身往往包含有随机性或不确定性. 这些性质既可能是系统的内在机制，也可能来源于系统中某些没有观测到的因素. 统计建模与统计学习方法是解决这类不确定性和复杂性问题的一种自然选择. 并且随着计算机和传感器技术的发展，使得直接收集大量的观测数据变得十分简单. 在缺乏基本原理模型时，可以利用这些观测数据在统计模型和统计学习的框架下来估计系统内部参数和其相关关系，最终可以从数据中得到需要的模型和系统的状态估计.

图模型的理论和应用包含非常丰富的内容，其统一的描述方法有助于设计通用、高效的学习算法，同时也使得在不同的研究领域和不同的研究方向中出现的新方法与新思路能在一定的高度得到交流和启发. 在应用图模型理论和相应的统计学习算法时，需要针对特定问题设计合适的模型结构. 建立在图模型框架下的统计学习方法是解决机器智能与模式识别研究中不确定性与复杂性问题的有效途径.

参 考 文 献

1 Alex Pentland. Looking at people: sensing for ubiquitous and sensable computing. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000, **22**(1):107～119

2 Christopher M Bishop. Neural Networks for Pattern Recognition. Oxford: Clarendon Press, 1995

3 Christopher M Bishop, Michel Jordan. Introduction to Graphical Model(Draft). To be appeared, 2002, Available at http://www-slab.usc.edu/courses/cs599-ATNN/Joudan-chapterz.pdf

4 Rosenblatt F. Principles of Neurodynamics. New York: Spartan, 1962

5 Widrow B, Michael A Lehr. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. In: Proc. IEEE, 1990, **78**(9):1415～1442

6 Steinbuch K. Die lernmatrix. Kybernetik. *Biological Cybernetics*, 1961, **1**:36～45

7 Rabiner L. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 1989, **77**(2):257～285

8 Vapnik V N, Chervonenkis A. On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk USSR*, 1968, **181**(4)

9 Tikhonov A N. On solving ill-posed problem and method of regularization. *Doklady Akademii Nauk USSR*, 1963, **153**:501～504

10 Parzen E. On estimation fo probability function and mode. *Annals of Mathematical statistics*, 1962, **33**(3):

11 Kolmogorov A N. Three approaches to the quantitative definitions of information. *Problem of Inform Transmission*, 1965, **1**(1):1～7

12 Minton Philips Johnston, Laird Hopfield. Neural networks and physical systems with emergent collective computational abilities. In: Proceedings of the National Academy of Sciences, 1982

13 Kohonen T. Self-Organizatıon and Associative Memory, 3rd edition. Berlin: Springer-Verlag, 1989

14 Kirkpatric S. Optimizatıon by simulated annealing: quantitative studies. *Journal Statıs. Phys.*, 1984, **34**:974～997

15 David H Ackley, Georey E Hinton, Terrence J Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 1985, **9**(1):147～169

16 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back—propagating errors. *Nature*, 1986, **323**:533～536

17 Kosko B. Fuzzy entropy and conditioning. Information Science,1986

18 Zhang Z H, Zheng N N, Wang T S. Fuzzy generalization of the counter-propagation neural network: a family of soft competitive basis function neural networks. *Soft Computing*, 2002,(1):

19 Zhang Zhi-Hua, Zheng Nan-Ning, Wang Tian-Shu. Fuzzy counter-propagation neural network, universal approximation, and application to time series prediction. In:International Conference on Neural Networks and Brain Proceedings, Beijing, 1998

20 Vladimir Cherkassky, Filip Mulier. Learning from Data: Concepts, Theory, and Methods. New York: Wiley, 1998

21 Vapnik V N. The Nature of Statistical Learning Theory, 2nd edition. New York: Springer-Verlag, 2000

22 Jensen F V. An Introduction to Bayesian Networks. UCL Press, 1996

23 Neal R M. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, University of Toronto, 1993

24 Michael I Jordan, Zoubin Ghahramanı, Tommi Jaakkola *et al*. An introduction to variational methods for graphical models. *Machine Learning*, 1999, **37**(2):183～233

25 张志华,郑南宁,王天树. 广义 LVQ 神经网络的性能分析及其改进. 自动化学报,1999,**25**(5):583～589

26 Comon P. Independent component analysis, a new concept? *Signal Processing*, 1994, **36**(3):287～314

27 Vladimır N Vapnik. The Nature of Statıstical Learning Theory. Heidelberg, DE:Springer-Verlag, 1995

28 Bernhard Scholkopf, Alex Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. Cambridge, MA:MIT Press, 1992

29 Dempster N M, Laird A P, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 1977, **39**:185～197

30 Gilks W, Richardson S, Spiegelhalter D. Markov Chain Monte Carlo in Practice. London: Chapman and Hall, 1996

31 Green P J. Reversible jump MCMC computation and Bayesian model determination. *Bıometrika*, 1995, **82**:711～732

32 Besag J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic systems. *Statistical Science*, 1995, **10**:3～66

33 Zhu S, Wu Y, Mumford D. Minimax entropy principle and its application to texture modeling. *Neural computation*, 1997, **9**:1527～1660

34 Yedidia J S, Freeman W T, Weiss Y. Generalized belief propagation. TR-2000-26, Mitsubishi Electronics Research Lab, 2000, Available at http://www.merl.com/papers/TR2000-26

35 Alexei A Efros, Thomas K Leung. Texture synthesis by non-parametric sampling. In: ICCV, 1999, (2):1033～1038

36 Alexei Efros, William Freeman. Image quilting for texture synthesis and transfer. In:Proc SIGGRAPH'01, 2001. 341～346

37 Lin Liang, Liu Ce, Xu Ying-Qıng *et al*. Real-time texture synthesis by patch-based sampling. MSR-TR-2001-40, Microsoft Research, 2001

38 Wang T S, Shum H Y, Xu Y Q, Zheng N N. Unsupervised analysis of human gestures. In: Lecture Note on Computer scıence 1385. ISBN:3-540-42680-9 Springer-verleg, 2001

39 Wang T S, Chen H, Zheng N N. An efficient object track algorithm in visual tracking system. In: Proceedıng of ASSC2000. ISBN 7-900033-85-8. 2000

40 Judea Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, California: Morgan Kaufman Publishers, 1988

41 Gregory F Cooper, Edward Herskovits. Abayesian method for the induction of probabilistic networks from data. *Machine Learning*, 1992, **9**:309～347

42 David Heckerman, Dan Geiger, David Maxwell Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In: Proc. KDD Workshop, 1994. 85～96

43 Kim J H, Pearl J. A computational model for causal and diagnostic reasoning in inference system. In: Proc. the 8th International Joint Conference on Artificial Intelligence, 1983. 190～193

44 Lauritzen S L, Spiegelhalter D J. Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society*, 1988, **50**:491～505

45 Shafer G, Shenoy P P. Probability propagation. *Annals of Mathematics and Artificial Intelligence*, 1990, **2**:327～352

46 Seneta E. Non-Negative Matrices and Markov Chains. New York: Springer-Verlag, 1981

47 Murphy K P. An ıntroduction to graphical models. Tech. Report, May, 2001, Available at http://www.cs.berkeley.edu/～murphyk/papers.html

48 茆诗松, 王静龙, 濮晓龙. 高等数理统计. 北京:高等教育出版社, 2000

49 McLachlan G J, Krishnan T. The EM Algorithm and Extensions. New York: Wiley, 1997

50 Lanterman A D. Schwarz, Wallace, and Rissane: intertwining themes in theories of model selection. *Int'l Statistical Review*, 2001, **69**:

51 Meng X L, Rubin D B. Recent extensions to the EM algorithm. In: Bayesian Statistics 4. Oxford University Press, 1992. 307～320

52 Celeux G, Forbes F, Mkhadri A. A component wise EM algorithm for mixtures, TR 3746, France, 1999. Available at http://www.inria.fr/RRRT/RR-3746.html

53 Ueda N, Nakano R. Deterministic annealing EM algorithm. *Neural Networks*, 1998, **11**:271～282

54 Ghahramani Z, Matthew J Beal. Graphical Models and Variational Methods. Advanced Mean Field Methods-Theory and Practive, MIT Press, 2000

55 Huber P J. Robust Statistics. New York: Wiley, 1981

56 Zoubin Ghahramani. Learning dynamic bayesian networks. In: Summer School on Neural Networks, 1997. 168～197

王天树 博士,IBM 北京研究院,研究领域为统计学习、计算机视觉与模式识别等.

郑南宁 中国工程院院士、教授,研究领域为智能信息处理、计算机视觉与模式识别等.

袁泽剑 博士研究生,研究领域为智能信息处理、计算机视觉与模式识别等.