概率模型与计算机视觉

概率模型与计算机视觉

林达华

美国麻省理工学院（MIT）博士

上世纪60年代, Marvin Minsky 在MIT让他的本科学生 Gerald Jay Sussman用一个暑假的时间完成一个有趣的Project: “link a camera to a computer and get the computer to describe what it saw”。从那时开始，特别是David Marr教授于1977年正式提出视觉计算理论，计算机视觉已经走过了四十多年的历史。可是，从今天看来，这个已入不惑之年的学科，依然显得如此年轻而朝气蓬勃。

在它几十年的发展历程中，多种流派的方法都曾各领风骚于一时。最近二十年中，计算机视觉发展最鲜明的特征就是机器学习与概率模型的广泛应用。在这里，我简单回顾一下对这个领域产生了重要影响的几个里程碑：

● 1984年：Stuart Geman和Donald Geman发表了一篇先驱性的论文：Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. 在这篇文章里，两位Geman先生引入了一系列对计算机视觉以后的发展具有深远影响的概念和方法：Markov Random Field (MRF)，Gibbs Sampling，以及Maximum a Posteriori estimate (MAP estimate)。这篇论文的意义是超前于时代的，它所建立的这一系列方法直到90年代中后期才开始被广泛关注。

● 1991年：Matthew Turk和Alex Pentland使用Eigenface进行人脸分类。从此，以矩阵的代数分解为基础的方法在视觉分析中被大量运用。其中有代表性的方法包括PCA, LDA，以及ICA。

● 1995年：Corinna Cortes和Vladimir Vapnik提出带有soft margin的Support Vector Machine (SVM)以及它的Kernel版本，并用它对手写数字进行分类。从此，SVM大受欢迎，并成为各种应用中的基准分类器。

● 1996年：Bruno Olshausen 和David Field 提出使用Overcomplete basis对图像进行稀疏编码(Sparse coding)。这个方向在初期的反响并不热烈。直到近些年，Compressed Sensing在信号处理领域成为炙手可热的方向。Sparse coding 在这一热潮的带动下，成为视觉领域一个活跃的研究方向。

● 90年代末：Graphical Model和Variational Inference逐步发展成熟。1998年，MIT出版社出版了由Michale Jordan主编的文集：Learning in Graphical Models。 这部书总结了那一时期关于Graphical Model的建模，分析和推断的主要成果——这些成果为Graphical Model在人工智能的各个领域的应用提供了方法论基础。进入21世纪，Graphical Model和Bayesian方法在视觉研究中的运用出现了井喷式的增长。

● 2001年：John Lafferty和Andrew McCallum等提出Conditional Random Field (CRF)。CRF为结构化的分类和预测提供了一种通用的工具。此后，语义结构开始被运用于视觉场景分析。

● 2003年：David Blei等提出Latent Dirichlet Allocation。2004年：Yee Whye Teh 等提出Hierarchical Dirichlet Process。各种参数化或者非参数化的Topic Model在此后不久被广泛用于语义层面的场景分析。

● 虽然Yahn Lecun等人在1993年已提出Convolutional Neural Network，但在vision中的应用效果一直欠佳。时至2006年，Geoffrey Hinton等人提出Deep Belief Network进行layer-wise的pretraining，应用效果取得突破性进展，其与之后Ruslan Salakhutdinov提出的Deep Boltzmann Machine重新点燃了视觉领域对于Neural Network和Boltzmann Machine的热情。

时间进入2013年，Probabilistic Graphical Model早已成为视觉领域中一种基本的建模工具。Probabilistic Graphical Model的研究涉及非常多的方面。 限于篇幅，在本文中，我只能简要介绍其中几个重要的方面，希望能为大家提供一些有用的参考。

Graphical Model的基本类型

基本的Graphical Model 可以大致分为两个类别：贝叶斯网络(Bayesian Network)和马尔可夫随机场(Markov Random Field)。它们的主要区别在于采用不同类型的图来表达变量之间的关系：贝叶斯网络采用有向无环图(Directed Acyclic Graph)来表达因果关系，马尔可夫随机场则采用无向图(Undirected Graph)来表达变量间的相互作用。这种结构上的区别导致了它们在建模和推断方面的一系列微妙的差异。一般来说，贝叶斯网络中每一个节点都对应于一个先验概率分布或者条件概率分布，因此整体的联合分布可以直接分解为所有单个节点所对应的分布的乘积。而对于马尔可夫场，由于变量之间没有明确的因果关系，它的联合概率分布通常会表达为一系列势函数（potential function）的乘积。通常情况下，这些乘积的积分并不等于1，因此，还要对其进行归一化才能形成一个有效的概率分布——这一点往往在实际应用中给参数估计造成非常大的困难。在变分推断中归一化是非常重要的操作。

值得一提的是，贝叶斯网络和马尔可夫随机场的分类主要是为了研究和学习的便利。在实际应用中所使用的模型在很多时候是它们的某种形式的结合。比如，一个马尔可夫随机场可以作为整体成为一个更大的贝叶斯网络的节点，又或者，多个贝叶斯网络可以通过马尔可夫随机场联系起来。这种混合型的模型提供了更丰富的表达结构，同时也会给模型的推断和估计带来新的挑战。

Graphical Model的新发展方向

在传统的Graphical Model的应用中，模型的设计者需要在设计阶段就固定整个模型的结构，比如它要使用哪些节点，它们相互之间如何关联等等。但是，在实际问题中，选择合适的模型结构往往是非常困难的——因为，我们在很多时候其实并不清楚数据的实际结构。为了解决这个问题，人们开始探索一种新的建立概率模型的方式——结构学习。在这

种方法中，模型的结构在设计的阶段并不完全固定。设计者通常只需要设定模型结构所需要遵循的约束，然后再从模型学习的过程中同时推断出模型的实际结构。

结构学习直到今天仍然是机器学习中一个极具挑战性的方向。结构学习并没有固定的形式，不同的研究者往往会采取不同的途径。比如，结构学习中一个非常重要的问题，就是如何去发现变量之间的内部关联。对于这个问题，人们提出了多种截然不同的方法：比如，你可以先建立一个完全图连接所有的变量，然后选择一个子图来描述它们的实际结构，又或者，你可以引入潜在节点(latent node)来建立变量之间的关联。

Probabilistic Graphical Model的另外一个重要的发展方向是非参数化。与传统的参数化方法不同，非参数化方法是一种更为灵活的建模方式——非参数化模型的大小（比如节点的数量）可以随着数据的变化而变化。一个典型的非参数化模型就是基于狄利克莱过程(Dirichlet Process)的混合模型。这种模型引入狄利克莱过程作为部件(component)参数的先验分布，从而允许混合体中可以有任意多个部件。这从根本上克服了传统的有限混合模型中的一个难题，就是确定部件的数量。在近几年的文章中，非参数化模型开始被用于特征学习。在这方面，比较有代表性的工作就是基于Hierarchical Beta Process来学习不定数量的特征。

基于Graphical Model 的统计推断 (Inference)

完成模型的设计之后，下一步就是通过一定的算法从数据中去估计模型的参数，或推断我们感兴趣的其它未知变量的值。在贝叶斯方法中，模型的参数也通常被视为变量，它们和普通的变量并没有根本的区别。因此，参数估计也可以被视为是统计推断的一种特例。

除了最简单的一些模型，统计推断在计算上是非常困难的。一般而言，确切推断(exact inference)的复杂度取决于模型的tree width。对于很多实际模型，这个复杂度可能随着问题规模增长而指数增长。于是，人们退而求其次，转而探索具有多项式复杂度的近似推断(approximate inference)方法。

主流的近似推断方法有三种：

(1)基于平均场逼近(mean field approximation)的variational inference。这种方法通常用于由Exponential family distribution所组成的贝叶斯网络。其基本思想就是引入一个computationally tractable的upper bound逼近原模型的log partition function，从而有效地降低了优化的复杂度。大家所熟悉的EM算法就属于这类型算法的一种特例。

(2)Belief propagation。这种方法最初由Judea Pearl提出用于树状结构的统计推断。后来人们直接把这种算法用于带环的模型（忽略掉它本来对树状结构的要求）——在很多情况下仍然取得不错的实际效果，这就是loop belief propagation。在进一步的探索的过程中，人们发现了它与Bethe approximation的关系，并由此逐步建立起了对loopy belief propagation的理论解释，以及刻画出它在各种设定下的收敛条件。值得一提的是，由于Judea Pearl对人工智能和因果关系推断方法上的根本性贡献，他在2011年获得了计算机科学领域的最高奖——图灵奖。

基于message passing的方法在最近十年有很多新的发展。Martin Wainwright在2003年提出Tree-reweighted message passing，这种方法采用mixture of trees来逼近任意的graphical model，并利用mixture coefficient和edge probability之间的对偶关系建立了一种新的message passing的方法。这种方法是对belief propagation的推广。

新的想法：当原来模型计算复杂的时候，建立一个对于某个模型的近似分解，然后用这个模型你逼近这个问题进而便于求解计算。

Jason Johnson等人在2005年建立的walk sum analysis为高斯马尔可夫随机场上的belief propagation提供了系统的分析方法。这种方法成功刻画了belief propagation在高斯场上的收敛条件，也是后来提出的多种改进型的belief propagation的理论依据。Thomas Minka在他PhD期间所建立的expectation propagation也是belief propagation的在一般Graphical Model上的重要推广。

(3)蒙特卡罗采样(Monte Carlo sampling)。与基于优化的方法不同，蒙特卡罗方法通过对概率模型的随机模拟运行来收集样本，然后通过收集到的样本来估计变量的统计特性（比如，均值）。采样方法有三个方面的重要优点。第一，它提供了一种有严谨数学基础的方法来逼近概率计算中经常出现的积分（积分计算的复杂度随着空间维度的提高呈几何增长）。第二，采样过程最终获得的是整个联合分布的样本集，而不仅仅是对某些参数或者变量值的最优估计。这个样本集近似地提供了对整个分布的更全面的刻画。比如，你可以计算任意两个变量的相关系数。第三，它的渐近特性通常可以被严格证明。对于复杂的模型，由variational inference或者belief propagation所获得的解一般并不能保证是对问题的全局最优解。在大部分情况下，甚至无法了解它和最优解的距离有多远。如果使用采样，只要时间足够长，是可以任意逼近真实的分布的。而且采样过程的复杂度往往较为容易获得理论上的保证。

蒙特卡罗方法本身也是现代统计学中一个非常重要的分支。对它的研究在过去几十年来一直非常活跃。在机器学习领域中，常见的采样方法包括Gibbs Sampling, Metropolis-Hasting Sampling (M-H)，Importance Sampling, Slice Sampling, 以及Hamiltonian Monte Carlo。其中，Gibbs Sampling由于可以纳入M-H方法中解释而通常被视为M-H的特例——虽然它们最初的motivation是不一样的。

Graphical Model以及与它相关的probabilistic inference是一个非常博大的领域，远非本文所能涵盖。在这篇文章中，我只能蜻蜓点水般地介绍了其中一些我较为熟悉的方面，希望能给在这方面有兴趣的朋友一点参考。

针对林达华老师的这篇综述，视觉计算研究论坛（以下简称「SIGVC BBS」：http://www.sigvc.org/bbs）提供了一个问答环节。林达华老师针对论坛师生提出的许多问题（如概率图模型与目前很热的深度神经网络的联系和区别）一一做了详细解答，如下所示。如果大家还有别的疑问，可继续提问，林老师有空会给予答复。

「SIGVC BBS」：最近深度学习受到机器学习和计算机视觉领域中研究人员的高度重视。然而，感觉有关深度学习一些理论并不是太完善。在计算机视觉领域，人们开始热衷于将其作为工具来使用。相对来讲，概率图模型已经有其完善的理论体系了。那么我们是

不是也可以完全用概率图模型这套理论来解释深度信念网络和深度Boltzman机?

林达华老师：从数学形式上说，Deep Network和Boltzmann machine可以看成是Graphical Model的特例。但是，目前在Graphical Model体系中所建立的方法，主要适用于分析结构较为简单的模型。而对于有多层latent layer的模型，现有的数学工具尚不能提供非常有效的分析。在NIPS 2012会议期间，我和Ruslan进行了交流。他们目前的主要工作方向还是进一步改善算法的性能（尤其是在大规模问题上的性能），以及推广这类模型的应用，尚未涉及深入的理论分析。

「SIGVC BBS」：基于Dirichlet过程的混合模型解决了确定组件数量的问题，这里面是否引入了其它的问题呢（比方说其它参数的确定）？除了不需要确定组件数量这一点之外，非参数化的模型还有其它哪些优势?

林达华老师：非参数化模型确实引入了其它参数，比如concentration parameter。但是，这个参数和component的个数在实用中是有着不同的影响的。concentration parameter主要传达的是使用者希望形成的聚类粒度。举个简单的例子，比如一组数据存在3个大类，每个大类中有3个相对靠近的子类。这种情况下，聚成3类或者9类都是合理的解。如果concentration parameter设得比较大，最后的结果可能形成9类，如果设得比较小，则可能形成3类。但是，如果人为地固定类数，则很可能导致不合理的结果。

需要强调的是非参数化贝叶斯方法是一个非常博大的方向，目前的研究只是处于起步阶段。而Dirichlet Process mixture model只是非参数方法的一个具体应用。事实上，DP像Gauss distribution一样，都是一种有着良好数学性质的过程（分布），但是它们在实用中都过于理想化了。目前的一个新的研究方向就是建立更为贴近实际的非参数化过程。相比于传统参数化方法而言，非参数化方法的主要优势是允许模型的结构在学习的过程中动态变化（而不仅仅是组件的数量），这种灵活性对于描述处于不断变化中的数据非常重要。当然，如何在更复杂的模型中应用非参数化方法是一个比较新的课题，有很多值得进一步探索的地方。

「SIGVC BBS」：文中后面提到的结构学习是不是这两年比较火的Structured Output Prediction呢？他们的关系如何？Structured Percepton和Structured SVM应该就是属于这个大类吗？结构学习的输出是树结构和图结构吗？结构学习与图像的层次分割或者层次聚类有关系吗?

林达华老师：Structured Prediction (e.g. Structured SVM) 其实属于利用结构，而不是我在文中所指结构学习。在大部分Structured Prediction的应用中，结构是预先固定的（比如哪些变量要用potential联系在一起），学习的过程其实只是优化待定的参数。尽管如此，这些工作本身是非常有价值的，在很多问题中都取得了不错的效果。

我在文中所提到的结构学习是指连结构本身都是不固定的，需要从数据中去学习。一般情况下，学习输出的是图或者树的结构（以及相关参数）。这个topic其实历史很长了，早期的代表性工作就是chow-liu tree。这是一种利用信息量计算寻找最优树结构来描述数据的算法。Alan Willsky的小组近几年在这个方向取得了很多进展。但是，总体而言这个方向

仍旧非常困难，大部分工作属于探索性的，并不特别成熟。目前在Vision中的应用不是特别广泛。但是，我相信，随着一些方法逐步成熟，进入实用阶段，它的应用前景是非常不错的。

「SIGVC BBS」：文中提到了Convolutional Deep Network、Deep Belief Network、Deep Boltzmann Machine等近年炙手可热的神经网络方法。那么，神经网络和概率图模型是不是本质上完全是一回事，只是观察角度和历史发展不同？感觉它们很多地方都很相似。深度学习里RBM学习的训练算法与概率图模型的学习推理算法有什么联系和区别吗？他们的结构模型有什么联系和区别吗？

林达华老师：这两类模型所使用的数学方法是非常不同的。Graphical model的很多推断和学习方法都有很深的数学根基。通过近十几年的努力，大家已经逐步建立起整套的方法论体系对相关算法进行分析。Deep Learning目前并没有什么有效的分析方法。Deep learning取得很好的性能，其中很多技巧性的方法(trick)起到了重要作用。至于为什么这些trick能导致更好的性能，目前还未能有一个很好的解释。

我个人看来，这些技巧其实是很有价值的：一方面，它们确实在实践中提高了性能；另外一方面，它们为理论上的探索提出了问题。但是，我觉得，有效回答这些问题需要新的数学工具（新的数学分析方法），这看来不是近期内能做到的。

「SIGVC BBS」：在一些论文中看到，采样的方法（如Gibbs采样）也有其缺点，一个是计算量比较大（computationally intensive），另一个是收敛检测比较难。不知道这些说法是否有道理，或者目前这些问题是否有得到解决？

林达华老师：这里提到的两个问题确实是Sampling的两个主要的困难。对于这些问题，过去几十年取得了很多进展，提出了很多新的采样方法，但是困难仍然很大。但是，采样能提供整个分布的信息，而且有渐近(asymptotic)的理论保证。这在很多情况下是一般的optimization方法做不到的。最近有新的研究尝试结合Sampling和Optimization，在特定问题上有一些有趣的结果——比如，George Papandreou的Perturb-and-MAP.

「SIGVC BBS」：在计算机视觉中，视觉目标跟踪问题已经用到了动态贝叶斯网络方法。一些最近发表的自然图像分割方法也用到LDA（Latent Dirichlet Allocation）。在受限的理想数据条件下，这些方法都取得了较好的结果。但是，不得不承认，我们在研究和应用的过程中，在心理上首先对应用概率图模型有所畏惧（这里除我们已经用得较多较熟悉的MRF、CRF和Dynamic Bayesian network based visual tracking—condensation之外）。主要的解释可能有：一方面，它不象很多正则化方法那样其细节能被自我掌握、观测和控制；另一方面，对于一个新的问题，我们需要不停地问自己：什么样的设计（图）是最好的。从而，在很多情况下，我们更愿意选择使用那些正则化方法。比如，对小规模人脸识别，我们会选择PCA＋LAD（SVM），对大一点的规模我们会考虑“特征选择＋adaboost”框架。就计算机视觉，能否从实践的角度给我们一点关于使用概率图模型的建议。另外，在计算机视觉中，什么样的问题更适合于采用概率图模型

方法来解决。

林达华老师：首先，Graphical model和其它的方法一样，只是一种数学工具。对于解决问题而言，最重要的是选择合适的工具，而不一定要选看上去高深的方法。对于普通的分类问题，传统的SVM, Boost仍不失为最有效的方法。

Graphical model通常应用在问题本身带有多个相互联系的变量的时候。这个时候Graphical model提供了一种表达方式让你去表达这些联系。我觉得并不必要去寻求最优的设计图，事实上，没有人知道什么样的图才是最优的。实践中，我们通常是根据问题本身建立一个能比较自然地表达问题结构的图，然后通过实验了验证这个图是不是合适的。如果不合适，可以根据结果分析原因对图做出修正。

举个具体的例子，比如对一个比赛视频进行分析。那么可能涉及多个变量：摄像机的角度，背景，运动员的动作等等。那么这个问题可能就设计多个未知变量的推断，这些变量间可能存在各种联系。这个时候，Graphical model可能就是一种合适的选择。

值得注意的是，选择合适的图有时候也需要一些经验。比如分布的选择上要注意形成conjugate，这样往往容易得到简易的推断公式。了解各种分布的特性以及它们可能对最后结果的影响也是有帮助的。