

人工神经网络的学习算法(2)

张鸿宾 (北京计算机学院)

上文我们分析了阶层型网络的学习算法。这里我们介绍另一种主要网络——相互结合型网络的学习算法。和阶层型网络不同,相互结合型网络的任意两个节点之间都可以有连接。信号可以在网络中传来传去。相互结合型网络和无反馈阶层型网络的形态如图1所示。在70年代,曾有人利用相互结合型网络开展联想存储器的工作^[1,2]。1982年以后,由于Hopfield等人在相互结合型网络上的一系列研究工作^[3,4,5],使得一度沉寂了的神经网络的研究又出现了新的热潮。相互结合型网络在解决约束满足、组合优化和联想存储等方面都有着很重要的应用。下面我们主要以Hopfield型相互结合网络以及它的一般化形式的Boltzmann machine为例,分析它们的基本性质和学习算法。

四、相互结合型网络——Hopfield网络, Boltzmann Machine和Gibbs Sampler

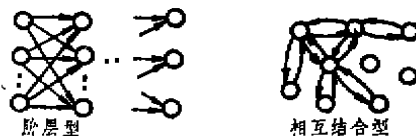


图1 阶层型和相互结合型网络的形态

Hopfield网络一般由二值节点组成。节点之间以对称的权连接,即 $W_{ij}=W_{ji}$ 。第 i 个节点的总输入信号为:

$$U_i = \sum_j W_{ij} \cdot S_j - \theta_i$$

式中 S_j 为节点 j 的状态或输出(0或1), θ_i 是节点 i 的阈值。

Hopfield网络假定节点状态的改变是非同步的,即每一时刻只有一个节点可以改变状态。

通过上面这些假定,我们可以在Hopfield网络上定义一个网络状态 $\alpha \in \{0, 1\}^n$ 的能量函数 $E(\alpha)$,

$$E(\alpha) = - \sum_{i,j} W_{ij} S_i S_j + \sum_i \theta_i \cdot S_i$$

如果网络状态的改变遵循下面的变化规则,即

IF $\sum W_{ij} S_j - \theta_i > 0$ THEN $U_i \rightarrow 1$

IF $\sum W_{ij} S_j - \theta_i < 0$ THEN $U_i \rightarrow 0$

IF $\sum W_{ij} S_j - \theta_i = 0$ THEN $U_i = U_i$

那么网络状态的改变总是朝着减小能量函数 $E(\alpha)$ 的方向进行,最后收敛到某个具有极小能量的平衡状态。需要注意的是,前面几节的误差函数是定义在连接权空间的,这里的能量函数则是定义在网络的状态空间上的。

一般地说,Hopfield网络的极小值点(平衡状态)有若干个。但这并不总是坏事,要看我们准备用Hopfield网络来干什么。当用Hopfield网络作联想存储器时,我们正是利用网络的各个极小值状态(平衡状态)来存储各个模式的。和现行数字机的番号地址型不同,这是一种状态地址或叫作内容地址型。对于联想存储器,我们需要解决模式的存储方法,分析网络的存储能力和抗噪声的能力等。

当我们用Hopfield网络求解组合最优化问题时,我们可以把最优化问题的评价函数和网络的能量函数对应起来,以此求出网络的各个连接权。然后让网络的状态不断向着减小能量函数,即评价函数的方向动作下去,最后收敛到能量函数的某个极值点(不一定是最小值点)。至于收敛到哪个极值点,则和网络的起始状态有关。实际应用问题要求我们尽量避开极小值点,最好找到最小值点。

为了解决上述问题,Hinton等人推广了Hopfield型网络^[7],在节点的状态变化规则

中引入了概率。他们把这种网络称为 Boltzmann Machine。

设 $\Delta E_i = E(S_i = 0) - E(S_i = 1)$, 那么节点 S_i 的状态由 0 变为 1 的概率 P_i 是由下式确定的:

$$P_i = \frac{1}{1 + \exp(-\Delta E_i/T)}$$

式中的 T 是称为“网络温度”的正实数。 P_i 的形状如图 2 所示。当 T 的值较大时, P_i 的形状趋于平坦。当 $T \rightarrow 0$ 时, P_i 的形状趋于阈值型。

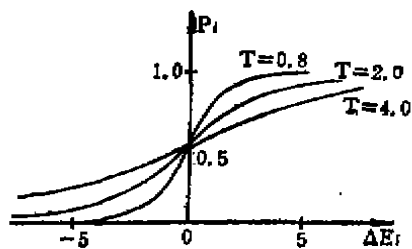


图2 $P_i = 1/[1 + \exp(-\Delta E_i/T)]$ 的图形

和 Hopfield 的变化规则不同, Boltzmann Machine 并不总是向减小能量的方向进行状态变化, 有时它的状态变化会增加状态的能量。而增加能量的状态变化的“容易程度”则和网络的“温度” T 有关。Boltzmann Machine 正是企图通过引进“噪声”的状态变化来跳出极小值点。这种思想来源于 Metropolis 等人的算法^[8]。

在一定的温度下, 从某个初始状态出发, 按照上述的状态变化规则变化下去, 网络可以达到某个“热平衡状态”。这是概率意义上的平衡。虽然在平衡状态时节点的状态仍在不断变化, 但状态的概率分布保持不变。这是马尔科夫过程的平衡分布。

假定有两个状态 α 和 β , 它们中间只有一个节点的状态不同, 只要一次状态变化就可以由 α 变为 β 。如果状态 α 和 β 的能量分别为 $E(\alpha)$ 和 $E(\beta)$ 的话, 那么根据上述的变化规则, 从状态 α 变到 β 的概率(转移概率) $P_{\alpha\beta}$ 为

$$P_{\alpha\beta} = \frac{1}{1 + \exp[-(E(\alpha) - E(\beta))/T]}$$

设状态 α 和 β 的个数分别为 n_α 和 n_β (这里不详细叙述 n_α 和 n_β 的准确含义了), 那么在平衡状态下有 $n_\alpha P_{\alpha\beta} = n_\beta P_{\beta\alpha}$ 。因此,

$$\frac{n_\alpha}{n_\beta} = \frac{P_{\beta\alpha}}{P_{\alpha\beta}} = \exp\left(\frac{E(\alpha) - E(\beta)}{T}\right)$$

由上式可以得到:

$$\frac{P_\alpha}{P_\beta} = \exp\left(\frac{E(\alpha) - E(\beta)}{T}\right)$$

$$P_\alpha = \frac{\exp(-E(\alpha)/T)}{\sum_\alpha \exp(-E(\alpha)/T)}$$

$$= C \cdot \exp(-E(\alpha)/T)$$

从上面两式可以看出, 网络的状态服从统计学上的 Boltzmann 分布, 两个状态出现的概率之比只和它们的能量差有关, 能量越小出现的概率越大。关于温度 T 的效果, 可以看出, T 越大则状态的变化越容易 (包括增加能量的变化), 陷入极小值的可能性就减小, 但同时停止在最小值的机会也随之减小。

根据以上的分析, 为了尽快地收敛到低温下的平衡状态, 我们可以从“高温”开始, 然后徐徐降温 (“退火”), 让网络的状态不断变化下去, 最后可以保证以相当高的概率收敛到网络的最小能量状态。这种方法叫作模拟退火法^[9]。

这里要注意的是如何安排降温的进程。如果降温的进程太快, 虽然可以较快地达到平衡状态, 但不能保证达到最小能量状态。从理论上讲^[10], 能够以概率 1 收敛到最小能量状态的充分条件是:

$$T(t) > T_0 / \log(t)$$

式中 $T(t)$ 是温度对时间的函数, T_0 是与网络的最大能量和最小能量之差有关的一个常数。由于实际运用时不可能如上式那样慢地进行, 所以只能说只要温度下降的相当慢的话, 就有相当大的可能达到能量最小点。

在 Boltzmann Machine 中, 每次只允许一个节点改变状态。如果能允许同时有几个

节点改变状态的话,那么网络能量“跳动”的幅度就大些。有些人基于这种思想曾作过理论上的分析和实际模拟。最近的一篇文章报告了使用Cauchy分布进行状态转移的模拟退火方法,其收敛到最小能量状态的充分条件是^[11]:

$$T(t) > T_0 / (1+t)$$

以上我们介绍了Boltzmann Machine的一些重要性质。从数学上看, Boltzmann Machine还可以推广到更一般的型式。S. Geman和D. Geman所提出的Gibbs Sampler就是其中的一个^[12]。他们的文章很长,公式和符号也很多。归纳一下后,其要点可以概括为:如果适当选择Boltzmann Machine状态变化规则中所使用的概率分布的话,那么平衡时的概率分布可以近似任意的分布。限于篇幅,这里就不准备详细分析了。

五、Boltzmann Machine的学习算法

Boltzmann Machine除了可以解决约束满足和组合优化问题之外,它还可以通过学习,模拟外界所给出的某种概率分布,实现联想存储。这一节我们介绍Boltzmann Machine的一种学习算法^[17]。

Boltzmann Machine的学习可以分为自我想起型和相互想起型两种情况(图3)。这两种网络的形态和功能虽然不同,但学习算法却相似。

自我想起型网络的节点分为V节点组和H节点组。其中V节点组作输入输出用。H节点组是隐蔽节点,没有也可以。设它们分别有k个和m个。学习的目的是,通过调整网络节点间的连接权,希望网络在自由动作时,能尽可能地逼近加在V上的外部模式的分布概率。学习可以按如下规则进行:

$$\Delta W_{ij} = \varepsilon (P_{ij} - P'_{ij}) \quad 0 < \varepsilon \leq 1$$

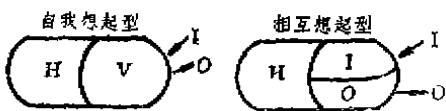


图3 自我想起型和相互想起型网络

式中的 P_{ij} 是整个网络完全自由动作、达到平衡状态时 $S_i S_j$ 的期望值。对于 P_{ij} ,假定加在V上的外部模式的概率分布为 $P(r)$, $r \in \{0, 1\}^k$ 。当把V节点固定在r状态,让H节点自由动作时,如果把 $S_i S_j$ 的期望记为 $P_{ij}(r)$,那么

$$P_{ij} = \sum_r P(r) \cdot P_{ij}(r)$$

在相互想起型网络中,网络分为输入节点组I,输出节点组O和隐蔽节点组H。设它们分别有k、l和m个。这种网络的学习规则和自我想起型的一样,只是 P_{ij} 和 P'_{ij} 的定义分别为:

$$P_{ij} = \sum_{\xi, r} P(\xi, r) \cdot E[S_i S_j; \xi, r]$$

$$P'_{ij} = \sum_{\xi} P(\xi) \cdot E[S_i S_j; \xi]$$

式中, $\xi \in \{0, 1\}^k$, $r \in \{0, 1\}^l$, 分别表示I节点组和O节点组的状态。 $P(\xi)$ 是把I节点组固定在外部的状态 ξ 时的概率。 $P(\xi, r)$ 是把I节点组固定在 ξ 状态、把O节点组固定在r状态(对于输入的 ξ 状态,希望的输出状态)时的联合概率。 $E[S_i S_j; \xi]$ 和 $E[S_i S_j; \xi, r]$ 分别为节点组固定在 ξ 状态或 (ξ, r) 状态,当达到平衡状态时 $S_i S_j$ 的期望值。

如何具体地实现上面 W_{ij} 的调整公式呢?这可以分两个阶段来进行。第一个阶段使 W_{ij} 和 P_{ij} 成比例的增加,第二个阶段使 W_{ij} 和 P'_{ij} 成比例的减小。两个阶段的持续时间相同,相互交替反复若干次。对于自我想起型网络,在第一阶段中,把从所希望的分布中随机选出的模式加在V节点组上(状态r)。然后在一定的时间间隔内,让别的节点自由动作起来。这时,如果有两个节点的状态同时为1,那么按一定的比例增加它们之间的连接权,亦即使 W_{ij} 和 $P_{ij}(r)$ 成比例地增加。然后,不断从所希望的分布中再选出其它的模式固定在V节点组上,重复上面的过程。在第二阶段中,让整个网络完全自由动作,如果有两个节点的状态同时为1,那么

按一定的比例减小它们之间的连接权。这样, W_{ij} 大致和 $P'_{ij}(r)$ 成比例的减小, 称为反学习。

对相互想起型网络, 在第一阶段中, 把从联合分布 $P(\xi, r) = P(\xi)P(r|\xi)$ 中随机选出的模式一个个地固定在 I 节点组和 O 节点组上, 使 W_{ij} 和 P_{ij} 成比例的增加。在第二阶段, 只把 I 节点组根据 $P(\xi)$ 固定在一个一个的输入状态进行反学习。要注意的是, 在相互想起型的学习过程中, 网络的 I 节点组总是固定在某个输入模式的状态。

有人把第一阶段称为觉醒状态, 把第二阶段称为睡眠状态。

下面我们分析一下上面学习算法的理论根据。

Boltzmann Machine 可以看作是一种外界概率分布的模拟机。它从加在输入节点组的输入模式中推测外界的概率结构, 然后加以学习和模拟。在自我想起的场合, 设加在输入节点组的概率分布为 $P(r)$, $r \in \{0, 1\}^n$ 。由 Boltzmann Machine 实际所实现的概率分布记为 $P'(r; W_{ij})$ 。我们可以采用统计学上的 Kullback 偏差 G 来表示这两个分布间的偏离:

$$G(W_{ij}) = D(P(r), P'(r; W_{ij})) = \sum_r P(r) \ln \frac{P(r)}{P'(r; W_{ij})}$$

$D(P, P') \geq 0$, 只有当 $P = P'$ 时, 等号才成立。

对于 G , Hinton 等人证明了下面的定理^[7]:

$$\text{定理1: } \frac{\partial G}{\partial W_{ij}} = -\frac{1}{T} (P_{ij} - P'_{ij})$$

如果没有隐蔽节点的话, 那么下面的定理 2 成立:

定理2: 如果没有隐蔽节点, G 的极小值只有一个。

对相互想起型的 Boltzmann Machine, 偏差 G 可以定义为:

$$G(W_{ij}) = \sum_{\xi} P(\xi) D(P(r|\xi), P(r|\xi; W_{ij}))$$

那么, 上面的两个定理仍然成立。

从上面的定理可以看出, 前面的学习规则正是使 G 极小的一种最速梯度下降法。

在实际应用上述学习算法时有许多计算问题需要考虑。例如, 学习时所给出的输入模式总是有限的, 那么输入节点组的其它状态出现的概率默认为 0。如果 Boltzmann Machine 的温度不为 0 的话, 这些状态的能量必须无穷大。因而连接权向无限大的方向发散。另一个问题是, 如果采用上面的最速下降法, 那么当能量函数的曲面呈峡谷状时, 会在两岸间振动起来。这是最速下降法的共同问题。有人采用了固定修正量的方法, 据报道取得了较好的效果。

Boltzmann Machine 的学习方法已经用到了声音识别系统上去, 取得了很好的效果。目前所面临的问题是, 在现行计算机上运行的时间太长。由于 Boltzmann Machine 的结构非常简单, 很容易硬件化。有些人通过利用模拟电路的热噪声来实现模拟退火, 试制了速度相当快的芯片。这项技术如果实用化, 必将大大促进人工神经网络的实际应用。

以上我们分析了人工神经网络几种主要的学习算法。除了上面介绍的之外, 还有许多别的算法。如竞争学习、协调学习、自适应学习等。正是这些学习算法使得人工神经网络具有了一定的学习能力和自我组织能力。正如程序设计使串行计算的原理得以普遍实现一样, 学习和自我组织很可能使并行计算得到普遍的应用。虽然还有许多理论问题需要进一步解决, 但学习能力、自我组织能力和并行计算能力无疑会成为下一代智能计算机的关键。

参考文献

- [1] T. Nagano, A model of Visual development, Biol. Cybernetics 26-1, 45-52, 1977.
- [2] T. Kohonen, E. Oja, Fast adap-

tive formation of orthogonalizing filters and associative memory in recurrent memory in recurrent networks of neuron-like elements, *Biol. Cybernetics* 21, 85—95, 1976.

- [3] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, *Proc. of the National Academy of Science USA* 79, 2254—2258, 1982.
- [4] J. J. Hopfield, Neurons with graded response have collective computational properties like those of two-state neurons, *Proc. of the National Academy of Science USA* 81, 3088—3092, 1984.
- [5] J. J. Hopfield, D. W. Tank, Neural computation of decisions in optimization problems, *Biol. Cybernetics* 62, 141—162, 1985.
- [6] G. E. Hinton, T. J. Sejnowski,

Optimal perceptual inference, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 448—453, 1983.

- [7] D. H. Ackley, G. E. Hinton et al, A learning algorithm for Boltzmann Machine, *Cognitive Sci.* 9, 147—169, 1985.
- [8] N. Metropolis, et al, Equation of state calculations for fast computing machines, *Journal of Chemical Physics* 6 1953.
- [9] S. Kirkpatrick, et al, Optimization by simulated annealing, *Science* 220, 671—680, 1983.
- [10] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images, *IEEE Trans. PAMI-6*, 721—741, 1984.
- [11] H. Szn, R. Hartley, Fast simulated annealing, *Physics Letters A*, Vol. 112, 157—162, 1987.

(上接第58页)

A_1, T_2), 如图7所示。

显然, 当用户需要时, 根据RDD可以自动地形成嵌套表的结构形式, 生成嵌套表NT (图8)。

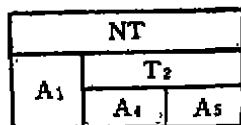


图8 嵌套表NT

嵌套表NT中的数据可以从关系 T_1 和 T_2 中得到。

用类似的方法, 可构造图形和屏幕的数据字典, 以实现相应的功能。

系统功能的扩充也是相当方便的, 如果要支撑CAD功能, 只需先以关系的形式存放点、线、面, 然后在其基础之上用相应的数据字典构造CAD的功能。

4. 结束语

从三级模式的一致性出发, 研制数据库系统的

软件集成, 不仅使得系统占内存少、功能强、用户界面友好, 数据冗余度小, 而且对于研究复杂对象的构造以及多介质数据库都是大有裨益的。

参考文献

- [1] D. S. Batory, T.Y. Leung and T.E. Wise, "Implementation Concepts for an Extensible Data Model and Data Language", *ACM Trans. on Database System*, Vol.13, No.3, 1988.
- [2] 田沧海、林钧海, "Fourth Generation Language NH₄GL—Towards an integration for software organization", *ICCC Symposium'89 on Computer Communication*, 1989.
- [3] 田沧海、郭红英、林钧海, "基于数据字典的用户友好接口", 小型微型计算机系统, Vol.9, No.9, 1988.