国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY
第16卷第4期 1994年12月 Vol. 16 No. 4

一种通用有效的神经网络映射算法*

王意洁 戴 葵 胡守仁

（国防科技大学计算机系 长沙 410073）

摘 要 首先分析了神经网络映射的本质，神经网络拓扑结构和神经计算过程，在充分考虑负载均衡与通讯开销的基础上，引入了时间步的思想，提出了一种通用有效的神经网络映射算法，最后给出的对多种拓扑结构的神经网络的测试结果证明了该算法的通用性和有效性。

关键词 神经网络映射算法；负载均衡；通讯开销

分类号 TP393

人工神经网络的研究主要分为三个方面：神经网络理论研究、神经网络应用研究和神经网络实现技术研究。神经网络的实现是用各种技术，合理有效地实现各种神经网络模型及其计算，它将神经网络理论研究和应用研究紧密联系起来。

在神经网络的虚拟实现中，并行神经计算机的研制是一个重要的具有广阔前景的研究方向。在神经计算机中，用P个物理处理单元（PE）去实现由N个神经元组成的神经网络，其中：P<N. P个物理处理单元（PE）之间需要分工协作，分工协作的目的是使多个处理单元能同时并发地执行任务，实现处理时间上的重叠，从而提高计算的速度。

如何分工，如何协作才能达到计算的高效，这就是映射算法所要完成的任务，也就是映射算法的功能。

研究映射算法的目的是：（1）极大地开发神经计算中的并行性；（2）处理单元之间负载均衡；（3）使系统通讯开销达到最小。

一般而言，映射算法与机器的体系结构、神经网络模型以及神经计算过程密切相关。并行电子神经计算机实现中所采用的传统映射方法主要有三种：（1）系统流水线处理方法；（2）样本批量处理方法；（3）层次交叉分割方法（Cross-layer）

上述三种方法是在不同层次上开发神经计算过程中潜在的并行性，并且适用于各自不同的神经计算机体系结构，从当前文献报道来看，对极大地开发神经计算并行性还没有统一有效的方法。所以我们试图能根据神经网络拓扑结构和神经计算本质特点来设计一通用有效的映射算法，进行神经网络的分割和处理机间的任务分配，使神经计算能在

* 863高科技项目，国家自然科学基金资助项目
1994年5月4日收稿

并行神经计算机上高效实现。

1 通用有效神经网络映射算法的理论基础

1.1 映射内容

神经网络的映射包含两个方面：①神经网络拓扑结构的映射；②神经计算过程的映射：神经计算的并行性与神经计算过程紧密相关，因而必须将神经计算过程与神经网络的拓扑结构相结合才能更好地开发神经计算的并行性，神经计算机中每个处理单元所要完成的计算主要有两类：

①神经元输入的加权求和，新的活跃值的计算，输出值的计算。

②神经元之间连接权值的修改。

将神经网络的计算看成是统一的信息流驱动模型，更能体现神经网络拓扑结构和神经计算过程的实质。这一模型与具体的神经计算过程相结合，便决定了神经计算中各种基本运算的顺序。实质上，神经计算是并行与串行相结合的一种计算。并行意味着每一时间步内可能有许多神经元与连接的计算可以同时进行，串行即表示某些神经元或连接的计算与其它一些神经元或连接的计算之间存在着顺序相关。那么根据神经计算过程和神经网络拓扑结构，我们总可以按神经计算的一个周期信息流驱动过程而对每一计算步内能同时进行并发计算的实体（神经元/连接）进行标记。即有：

计算步 t_i：u_1，u_2，…，u_m 能并行计算（u_i 表示神经元）；

计算步 t_i：c_1，c_2，…，c_n 能并行修改（c_i 表示连接）。

我们由以上讨论可以得到如下定理：

定理 1 在神经计算的一个周期内，总可以根据神经计算过程和神经网络模型的拓扑结构对神经元/连接的各种类型计算进行时间标记，具有相同标记的计算可以并行执行。

定理 2 一个周期内的同一时刻不可能出现同一实体（神经元/连接）有两种不同类型的计算。

定理 3 具有较小标记时刻的计算必须在具有较大标记时刻的计算之前完成。

定理 4 某一实体同一类型计算的周期为 T.

定理 5 t_{i+1}时刻进行计算的实体必与在 t_i 时刻进行计算的实体存在相关。

1.2 映射原则

为了取得良好的映射效果，保证神经计算在 并行电子神经计算机上有效实现，映射分配是依据由神经网络拓扑结构和神经计算过程得到的并行计算模型进行的，并适当考虑机器的体系结构。

对神经网络模型中的神经元进行映射分配主要依据如下映射原则：

①在通讯开销能够容忍的前提下，应尽可能地保证处理单元间负载均衡。

②在映射分配过程中，神经元是根据时间步一组组进行分配的，每一组神经元都是在综合考虑其对负载均衡和通讯开销的影响后才进行分配的。

1.3 负载均衡和通讯开销

对负载均衡和通讯开销两个关键因素衡量的好坏将直接影响映射分配的效果。在分

析讨论神经网络拓扑结构、神经计算过程和机器体系结构的基础上，我们给出了在映射分配过程中衡量负载均衡和通讯开销的方法。

通讯开销：通讯开销不仅与不同处理单元上的神经元之间的连接数有关，还与处理单元间的通讯方式有关。基于上述考虑，我们给出改进后的通讯开销衡量公式：

$$\gamma = \frac{T_{comm}^{sum}}{C} = \frac{\sum_{i=1}^{N} T_{comm}(i)}{C} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{N} O_{ij}T_{ij})}{C}$$

其中：γ 为通讯开销因子；C 为神经网络模型中神经元间连接数；O_{ij} 为处理单元 i 中的神经元与处理单元 j 中的神经元之间的连接数；T_{ij} 为处理单元 i 与处理单元 j 之间的通讯方式因子；N 为处理单元个数。

负载均衡：负载均衡是指并行执行任务的多个处理单元之间的任务分配是否均衡，也就是每个时刻各处理单元是否任务饱满且均衡。神经计算是并行与串行相结合的计算过程，每个计算步中多个神经元并行计算、多个计算步串行构成整个计算过程。并行计算模型中的每一组神经元即是神经计算过程中的每个计算步内能同时进行并发计算的神经元。整个映射分配的负载均衡好坏是由每组神经元分配时的负载均衡好坏共同决定的。改进的负载均衡衡量公式如下：

$$K_{all} = \frac{1}{\text{steps}} \sum_{i=1}^{\text{steps}} K_i$$

其中：K_{all} 为映射分配后总的负载不均因子；steps 为映射分配的神经元组的个数；K_i 为分配第 i 组神经元时的负载不均因子。

$$K_i = \frac{\sum_{j=1}^{N} |T\text{calc}(j) - T\text{ave/calc}|}{N \times T\text{ave/calc}}$$

N 为处理单元个数；$T\text{ave/calc} = \frac{1}{N}\sum_{j=1}^{N} T\text{calc}(j)$；$T\text{calc}(j)$ 为第 j 个处理单元的计算时间。

2 通用有效的神经网络映射算法

2.1 算法描述

基于对映射算法理论上的探讨，可以有如下的通用有效神经网络映射算法：

STEP-1：读入神经网络和处理单元的有关信息（如神经元数目、处理单元数目等）。

STEP-2：根据神经网络的拓扑结构和神经计算过程形成并发计算模型，并由此对神经元进行分组。

STEP-3：初始化。

STEP-4：$t=1$.

STEP-5：对时刻 1 能并发计算的神经元进行映射分配。

STEP-6：若 $t<$steps（steps 为神经元组的数目），则 $t=t+1$；否则转 STEP-9.

STEP-7：设时刻 $(t-1)$ 能并发计算的神经元为 u_1、u_2、…、u_k，时刻 t 能并发计算的神经元为 v_1、v_2、…、v_m，由时刻 $(t-1)$ 和时刻 t 能并发计算的神经元形成连接矩阵如下：（见图 1）

STEP-8：对图 1 中的连接矩阵进行行变换，若能满足映射分配对负载均衡和通讯开销的要求则分配神经元到各处理单元，并转 STEP-6. 否则，对连接矩阵进行行列变换，如果仍不能满足映射分配的要求，则只要求行变换到负载不均因子和通讯开销因子都为当前能得到的最小值为止，将神经元分配给各处理单元，转 STEP-6；如果能满足映射分配的要求，则按连接矩阵改变 $t-1$ 时刻的神经元分配，若这种改变能满足 $t-1$ 时刻映射分配的要求则分配 t 时刻神经元到各处理单元，并转 STEP-6，否则依次调整 $t-2$ 时刻神经元的分配，$t-3$ 时刻神经元的分配……直到哪一次满足了要求则分配 t 时刻神经元并转 STEP-6，在反向调整过程中，若不能满足映射分配要求则取通讯开销因子和负载不均因子当前能得到的最小值，然后继续向前调整，若调整完第一层仍不能满足要求，则取当前通讯开销因子和负载不均因子的最小值，并分配 t 时刻神经元，转 STEP-6.

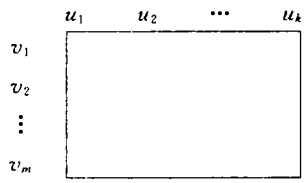

图 1 相关矩阵

STEP-9：输出分配结果，得出整个映射分配的负载不均因子 K_{all} 和通讯开销因子 γ.

STEP-10：算法结束。

2.2 模拟结果

我们已利用该算法对多种神经网络模型进行映射分配，实验结果证明了该算法是通用有效的，这里我们给出一个验证实例及其验证结果。

神经网络模型：

神经网络：10 处理单元：4 神经网络：3

每层的神经元情况：

第 1 层－1 2 3 第 2 层－4 5 6 7 第 3 层－8 9 10

神经网络层与层之间的连接情况：

第 1 层－第 2 层全互连 第 2 层－第 3 层全互连

处理单元通讯方式：

广 播 式

分配结果：

PE1： 1 4 8 PE2： 2 5 9 PE3： 3 6 10 PE4： 7

通讯开销：γ==0.75 负载均衡：K_{all}=0.333

3 讨　论

从分析神经网络映射的本质出发，综合神经网络拓扑结构和神经计算过程作为映射分配的依据，设计了通用有效的神经网络映射算法。该算法在理论上的突破是时间步的引入，这也是对神经计算过程的本质发现，它为在映射分配过程中真正地实现负载均衡和减少通讯开销起了决定性作用。从实验结果来看，效果很好，这种自动的任务分配大大减轻了用户使用系统的负担，并且算法适用于多种神经网络模型，由于该算法充分考虑了减少通讯开销和负载均衡，保证了神经计算的有效实现。

参 考 文 献

1 hunekuan Chcng. The Optimal Patitioniog of Netwerks. Networks. 1992, 22: 297～315

2 戴葵，刘燕，王意洁，张春元，胡守仁．神经网络的一种通用层次交叉分割映射算法．第三届中国神经网络大会论文，西安，1993

3 Dai kui, Yijie Wang, Shouren Hu. Some Applications of Object Oriented Methodology in Neurocomputer System Dcsign. Inteinational Conference on Neural Network and Signal Processing (ICNNSP' 93)

4 胡守仁，余少波，戴葵．神经网络导论．国防科技大学出版社，1993

5 Philipp Schmid. The Mapping Problem. A Ncural Nciwork Approach. International Neural Network Conference, Paris, France, 1990

6 王意洁，戴葵，张春元，胡守仁．计算机中几种神经网络信息存储方法及其性能评价．第三届中国神经网络大会论文集，西安．1993

A General-purpose Efficient Neural Network Mapping Algorithm

Wang Yijie　Dai Kui　Hu Shouren

(Department of Computer, NUDT, Changsha, 410073)

Abstract

Neural network mapping algorithm is an important research content in the field of neural netork virtual implementation. The efficiency of the algorithm affects the virtual implementation of neural network. In this paper, the nature of neural network mapping, the neural nctwork topology and the neurocomputing process have been analysised at first. Based on thorovgh consideration of load-balance and communication-cost, the thesis gives a general-purpose efficient neural network mapping algorithm. introducing the idea of timestep. Finally, the experimenting results of neural networks with different topologies are presented, which proves the generality and efficiency of the algorithm.

Key words　neural network mapping algorithm; load-balance; communication-cost