

8) KBANN 是根据由试验得出的经验性参数, 计算、设置初始权值和阈值, 缺乏理论上的分析和依据。

三、结束语

从以上的发展历史, 我们发现有以下几个方面的工作还需进一步的研究:

1. 神经网络的学习包括两方面, 权值的变化和拓扑结构的改变。目前的神经网络学习算法主要集中于权值的变化上, 利用符号机器学习技术辅助神经网络学习, 加快权值学习过程, 另外使网络拓扑结构能合理地确定和变化, 形成神经网络的一个统一学习算法(权值和拓扑结构统一变化)。目前还没有一个通用的、独立于问题的方法来选择一个好的网络拓扑结构, 形成合理的网络拓扑结构是目前研究的一个热点。

2. 利用符号学习技术确定神经网络何时停止训练, 防止训练过度或训练不足。

3. 如何利用符号学习技术减少学习例子, 如用符号规则指导神经网络尽快结束训练, 因为有时学习例子很难找到或寻找代价很大。

4. 训练例子的顺序与学习结果的关系问题。

5. 较好地处理带有噪音的数据以及个别有错误的数据。

6. 从神经网络中抽取符号规则是联接机制与符号机制学习相互转化的一种途径, 也是两种机制结合的一个方法。同时使神经网络具有解释能力, 便于更好地理解神经网络, 目前也是一个研究热点。

7. 由于网络中联接权与知识库中规则强度有一定的对应关系, 联接权的修正对应于规则强度的修正, 从而为规则强度的学习提供了有效的手段。

8. 一般神经网络训练前, 随机设置初始权值, 然后开始训练, 研究表明[Towell et al. 1994]: 初始值影响神经网络学习, 如何利用符号规则选择一个好的初始点(初始权值, 初始网络拓扑结构, 参数)是需要深入研究的。

9. 符号学习中归纳学习与神经网络学习相结合问题。这两者有共同之处, 神经网络是从例子中学习, 也是一种归纳学习。

10. 解释学习与神经网络相结合, 发展基于解释的神经网络(EBL-ANN), KBANN 就是 EBL-ANN 算法的发展。

11. 进一步提高神经网络的可靠性、鲁棒性。

总之, 联接机制与符号机制相结合的机器学习是一个值得研究的方向, 存在许多问题, 研究下去, 在机器学习或神经网络学习领域一定会有所突破。(参考文献共 38 篇略)

(上接第 62 页)

根据预备定理 4, 第 i 小组的数据一致趋向于均匀分布($1 \leq i \leq N$)。□

定理 2 递归分组排序的期望复杂度为 $O(N)$ 。

证明: 由于递归分组排序算法第一次分组的期望时间为 $O(N)$; 又 第一次分组后, 再对每个小组分组排序, 即用桶排序算法排序每个小组的数据, 由定理 1, 第一次分组后, 每个小组内的数据服从均匀分布, 故排序第 r 个小组数据的平均工作量小于 CN_r , 其中 N_r 是第 r 个小组的数据个数, C 为一常数, $1 \leq r \leq N+1$ 。于是, 排序全部 $N+1$ 个小组的数据的平均工作量小于:

$$\sum_{r=1}^{N+1} C \cdot N_r = CN$$

故 递归分组排序的期望时间为 $O(N)$ 。□

递归分组排序不要求数据服从某些特殊概率分

布, 其排序平均工作量为 $O(N)$ 。优于 Akl 等人出的桶排序算法。

4 实验结果

在 IBM-PC 机上, 采用 BASIC 语言, 对同一组随机数(由 RND 函数产生), 分别使用 Quicksort 和递归分组排序算法, 所用的 CPU 时间(单位为秒)如表 1 所示。

表 1

数据个数	200	500	1000	1500
Quicksort(秒)	31	84	185	34
递归分组排序	18	44	87	130

实验结果表明递归分组排序算法的排序速度比快速排序算法的排序速度要快一倍, 且数据量愈大, 递归分组排序愈优。(参考文献共 5 篇略)

一个期望时间为 $O(N)$ 的递归分组排序

徐绪松

周建钦

(武汉大学管理学院) (曲阜师范大学数学系)

摘要 This paper provides a sort algorithm, using a formula to recursive divided elements into groups. This algorithm sorts a sequence of n elements in $O(n)$ expected time, when the elements obey general probability distribution.

关键词 Sort, Recursive grouping.

1982年, Akl 等人提出桶排序算法^[1], 若排序文件中的数据服从概率分布, 其密度函数有界, 则桶排序的平均工作量为 $O(N)$, 特殊的, 若排序文件中的数据服从均匀分布, 则桶排序的平均工作量也为 $O(N)$. 但当数据服从正态分布时, 由于数据范围无界, 桶排序的平均工作量大于 $O(N)$.

本文提出递归分组排序, 当数据服从一般概率分布时, 算法排序 N 个数据元素的期望时间为 $O(N)$.

1 递归分组排序

输入 数据文件 $F: y(i) \quad 1 \leq i \leq N$

开辟 N 个工作单元 $w()$, 递归分组排序算法描述如下:

1) $w(i) \leftarrow y(i) \quad 1 \leq i \leq N$

2) 分别求 $w()$ 的最大值 $\max$, 最小值 $\min$, 若 $\max = \min$, 则结束; 否则

3) 若 $J = \lfloor (w(i) - \min) * K / (\max - \min) \rfloor + 1$, 则将 $w(i)$ 分入第 J 组, $1 \leq i \leq N, K = N$. 将 $w()$ 中的数据放回原来的存储单元中, 组号小的数据放在前面, 同一组内的数据连续存放.

4) 依次检查每个小组内的数据个数 r , 若 $r > 1$, 则令 $K = r$, 将这一小组数据存放在 $w()$ 中, 做 2—4, 否则

5) 结束.

2 几个预备定理

预备定理1 (Borel-Cantell 引理)^[2] 若随机事件序列 $\{A_n\}$ 满足 $\sum_{n=1}^{\infty} P(A_n) < \infty$, 则 $P\{\lim_{n \rightarrow \infty} \bar{A}_n\} = 1$.

预备定理2 若 y 为一随机变量, 则 $E(y) < \infty$

成立的充要条件为 $\sum_{n=1}^{\infty} P(|y| > n) < \infty$ ^[4].

预备定理3 若 $y(1), y(2), \dots, y(N)$ 为独立同分布随机变量, 且 $E|y(1)| < \infty$, 则 $\lim_{N \rightarrow \infty} (\max - \min)/N = 0$. 其中 $\min = \min\{y(1), y(2), \dots, y(N)\}$, $\max = \max\{y(1), y(2), \dots, y(N)\}$

证明: (提示) 根据预备定理2和1可证.

预备定理4 若随机变量 x 的密度函数为 $f(x)$, $f(x)$ 的最大值为 D , x 的取值区域为 $[b_1, b_2]$, 令 $B = b_2 - b_1$, 则随机变量 x 为均匀分布的充要条件是 $D \cdot B = 1$.

证明 (略).

3 算法分析

定理1 若数据文件 F 中的数据服从一般概率分布, 其密度函数 $f(y)$ 连续, 则在递归分组排序中, 经第一次分组后, 每小组的数据一致趋向于均匀分布.

证明 文件 F 中的数据密度函数为 $f(y)$, 有第 i ($1 \leq i \leq N$) 小组数据的分布密度为:

$$f(y) / \int_{\min + (i-1)w}^{\min + iw} f(t) dt$$

其中 w 为小组宽度, $w = (\max - \min)/N$.

设 $f(y)$ 在第 i 个小区间上的最大值为 $f(y_1)$, 由积分中值定理并根据密度函数的性质, 由于 $f(y)$ 连续, 易证 $f(y)$ 一致连续.

根据预备定理3, 有 $f(y_1)/f(y_0)$ 一致趋向于1 (其中 y_1, y_0 均属第 i 个小区间). 由于第 i 小组数据的

概率分布最大值为 $f(y_1)/f(y_0)w$, 所以 $w \cdot \frac{f(y_1)}{f(y_0)w} = \frac{f(y_1)}{f(y_0)}$ 一致趋向于1. (下转第9页)