

基于约束规划的选择性神经网络集成方法

龚安 张敏

(中国石油大学(华东)计算机与通信工程学院 东营 257061)

摘要 提出了一种基于约束规划的选择性神经网络集成方法,在训练出个体网络之后,用约束规划方法选择出相对最佳的个体网络组成神经网络集成。理论分析和实验结果表明,该方法设计过程简单,能够以较小的运算代价提高神经网络集成的泛化能力。

关键词 神经网络,神经网络集成,约束规划

Constrained Programming Based Selective Neural Network Ensemble Method

GONG An ZHANG Min

(School of Computer Science and Communication Engineering, China University of Petroleum, Dongying 257061)

Abstract A selective constructing approach to neural network ensemble based on constrained programming is presented. Constrained programming is used to select the trained individual networks to construct a new ensemble. Theoretical analyses and experimental results show that the design of this method is simple and it improves the generalization ability of neural network ensemble at low cost.

Keywords Neural network, Neural network ensemble, Constrained programming

1 引言

随着人工神经网络理论的逐渐成熟、高速计算与并行处理的发展,神经网络集成已成为机器学习和神经计算领域的新的研究热点,也成为人工神经网络研究向广度和纵深发展的一种趋势。

Krogh 等人^[1]通过理论分析得出,组成神经网络集成的各网络精度越高、差异度越大,集成的效果越好。为了提高集成的效果,一些研究者^[2,3]提出了选择性集成的方法,即独立训练生成若干个神经网络之后,不对所有的网络进行集成,而是选择出相对最佳的个体网络组成集成。

目前常用的选择性集成方法有简单枚举法、贪心算法和遗传算法等。简单枚举法可以获得最优解,但在个体网络较多时,该方法由于计算复杂度太大而无法实现。基于贪心法的选择性集成计算复杂度较小,但是容易陷入“局部极小点”,通常情况下不能得到最优解,即在验证集上的泛化误差较大且不稳定,这使得集成的精度不高且可靠性差。遗传算法计算效率比较低,使得基于遗传算法的选择性集成计算复杂度更大。

针对以上问题,本文提出一种基于约束规划的选择性神经网络集成方法。该方法一般情况下能够得到最优解,且与遗传算法相比,少了对大量参数的选择,效率较高,能以较小的代价提高神经网络集成的泛化能力。

2 理论分析及模型建立

对于神经网络集成的结论生成方法为简单平均法的情况,文[3]经过推导得出了以下结论:在剔除满足一定条件的神经网络后,集成的泛化能力会得到提高。本文对结论采用加权平均法的神经网络集成,去除部分个体网络后的情况进行讨论。

假设神经网络集成由 N 个神经网络 $f_1, f_2, \dots, f_N$ 组成,

各网络分别被赋予权值 $w_i (i=1, 2, \dots, N)$, 满足 $\sum_{i=1}^N w_i = 1$ 且 $0 \leq w_i \leq 1 (i=1, 2, \dots, N)$ 。集成采用加权平均法对结论进行合成。

假设 $x \in R^m$ 满足分布 $p(x)$, 在输入 x 下对应的期望输出为 $d(x)$, 个体网络 $f_i (i=1, 2, \dots, N)$ 的实际输出为 $f_i(x)$, 则神经网络集成在输入 x 上的实际输出为:

$$\bar{f}(x) = \sum_{i=1}^N w_i f_i(x) \quad (1)$$

周志华教授等^[4]推导出神经网络集成的泛化误差的公式如下:

$$E = \sum_{i=1}^N \sum_{j=1}^N w_i w_j C_{ij} \quad (2)$$

其中 C_{ij} 为个体网络 f_i 与 f_j 的相关度, 定义为:

$$C_{ij} = \int dx p(x) (f_i(x) - d(x))(f_j(x) - d(x)) \quad (3)$$

且 C_{ij} 满足

$$C_{ii} = E_i \text{ 且 } C_{ij} = C_{ji}$$

由式(2), 考虑从神经网络集成中去除个体网络 f_k 后, 由剩余的 $N-1$ 个神经网络 $f_1, f_2, \dots, f_{k-1}, f_{k+1}, \dots, f_N$ 组成集成, 此时, 各网络的权值发生了变化, 记为 $w'_i (i=1, 2, \dots, k-1, k+1, \dots, N)$, 满足且 $\sum_{i=1, i \neq k}^N w'_i = 1$ 且 $0 \leq w'_i \leq 1 (i=1, 2, \dots, k-1, k+1, \dots, N)$ 。去除个体网络 f_k 后的神经网络集成的泛化误差为:

$$E' = \sum_{i=1, i \neq k}^N \sum_{j=1, j \neq k}^N w'_i w'_j C_{ij} \quad (4)$$

则有

$$E - E' = \sum_{i=1, i \neq k}^N \sum_{j=1, j \neq k}^N (w_i w_j - w'_i w'_j) C_{ij} + 2 \sum_{i=1}^N w_i w_k C_{ik} + w_k^2 C_{kk} \quad (5)$$

在式(6)成立时, $E > E'$, 即去除个体网络 f_k 后神经网络集成的泛化能力得到了提高。

$$\sum_{i=1}^N \sum_{j \neq k}^N (w'_i w'_j - w_i w_j) C_{ij} < 2 \sum_{i=1}^N w_i w_k C_{ik} + u_k^2 E_k \quad (6)$$

理论分析表明,对于结论生成方法采用加权平均法的神经网络集成,在尽可能独立地训练出多个神经网络之后,不一定要把所有的网络全部集成起来。如果有某个网络满足式(6),则在将其去除后,选择其他网络进行集成将获得更强的泛化能力。

神经网络集成中各个体网络的权系数反映了该个体网络在集成中的重要性,因此,可以求出个体网络的最优权系数,去除权系数小于预设阈值的网络,依此对个体网络进行选择。关键的问题是如何求得最优权系数。由于神经网络集成的泛化误差是各个体网络的权系数的函数,而最优权系数则是使得神经网络集成的泛化误差达到最小的各网络的权系数,因此,本文提出建立约束规划的数学模型来解决这个问题,目标函数是神经网络集成的泛化误差 E 最小,约束条件是个体网络的权系数之和等于 1,并且每个权系数都大于 0 小于 1。模型建立如下:

$$\begin{aligned} \min \phi(w) &= E = \sum_{i=1}^N \sum_{j=1}^N w_i w_j C_{ij} = w C w^T \\ \text{s. t. } &\sum_{i=1}^N w_i = 1 \\ &0 \leq w_i \leq 1 (i=1, 2, \dots, N) \end{aligned}$$

基于约束规划的选择性神经网络集成方法的基本思想是:在独立训练出一批神经网络之后,利用约束规划方法来选择相对最佳的个体网络集合,从而使构造出的神经网络集成具有更好的泛化能力。

3 算法描述

由于在解决实际问题时,无法求得个体网络乃至整个集成的泛化误差的精确值,通常使用 Holdout 方法^[3]实现神经网络集成中个体神经网络的选择。以神经网络集成在验证集上的平均误差作为其泛化误差的估计值。假设验证集为 V ,则上文中的约束规划数学模型变为如下形式:

$$\begin{aligned} \min \phi(w) &= E_w^V = \sum_{i=1}^N \sum_{j=1}^N w_i w_j C_{ij}^V = w C^V w^T \\ \text{s. t. } &\sum_{i=1}^N w_i = 1 \\ &0 \leq w_i \leq 1 (i=1, 2, \dots, N) \end{aligned}$$

其中, E_w^V 为与 w 对应的神经网络集成在验证集 V 上的平均误差。个体网络 $f_i (i=1, 2, \dots, N)$ 所对应的权值为 $w_i (i=1, 2, \dots, N)$, 令 $w = (w_1, w_2, \dots, w_N)$ 。

基于约束规划的选择性神经网络集成的算法描述如下:

a) 在初始训练集上通过 Bootstrap 方法产生 N 个子训练集,用这 N 个子训练集独立训练出 N 个神经网络,组成集合 $\{f_1, f_2, \dots, f_N\}$;

b) 分别计算在验证集 V 上的个体网络 f_i 和 f_j 的相关度估计值:

$$C_{ij}^V = \frac{1}{V} \sum_{x \in V} (f_i(x) - d(x))(f_j(x) - d(x))$$

$i, j = (1, \dots, N);$

c) 求解约束规划问题:

$$\begin{aligned} \min \phi(w) &= E_w^V = \sum_{i=1}^N \sum_{j=1}^N w_i w_j C_{ij}^V = w C^V w^T \\ \text{s. t. } &\sum_{i=1}^N w_i = 1 \\ &0 \leq w_i \leq 1 (i=1, 2, \dots, N) \end{aligned}$$

得其最优解 $w^* = (w_1^*, w_2^*, \dots, w_N^*)^T$;

d) 去除 w^* 中小于预设阈值的分量对应的个体网络,这样得到子集合 $\{f_1, f_2, \dots, f_s\}$, 与之对应得到选择性神经网络

集成的输出为 $f_{\text{final}} = \sum_{i=1}^S w'_i f_i, w'_i$ 的值由集成采用的具体的结论生成方法决定,本文采用加权平均法。

4 实验结果及分析

本文用两个数据集——Friedman #3 和 Mexican Hat 对本文方法和传统方法进行实验比较。对每个数据集,首先独立地训练出 20 个 BP 网络(1 个隐含层),然后用本文方法选择出部分神经网络组成集成。对于传统方法,即使用全部 20 个神经网络组成集成,对每个数据集,都用上述两种方法进行 10 次实验,分别记录它们在 10 次实验中产生的神经网络集成的测试集误差平均值以及误差的修正标准差,结果如表 1 所示。实验中,本文方法中的约束规划方法的实现采用了 Matlab 的 Optimization 工具箱,预设阈值取 0.05。

第 1 个数据集(Friedman #3 数据集)每个输入向量含有 4 个分量,数据集根据下式产生:

$$y = \tan^{-1} \frac{x_2 x_3 - \frac{1}{x_2 x_4}}{x_1} + \epsilon$$

其中 x_1 满足 $[0, 100]$ 上的均匀分布, x_2 满足 $[40\pi, 560\pi]$ 上的均匀分布, x_3 满足 $[0, 1]$ 上的均匀分布, x_4 满足 $[1, 11]$ 上的均匀分布, ϵ 服从正态分布 $N(0, \sigma^2)$, σ 取 0.1, 使得信噪比为 3:1。用上述数据集产生 1000 个示例,训练集大小为 400, 以其中的 200 个示例作为验证集,测试集大小为 600。实验结果如表 1 所示。

第 2 个数据集(Mexican Hat 数据集)根据二维墨西哥草帽函数生成,如下式所示:

$$t = \frac{10 \sin x}{x} + \epsilon$$

其中 x 满足 $[-\pi, \pi]$ 上的均匀分布,服从正态分布 $N(0, 1)$ 。用上述数据集产生 480 个示例,训练集大小为 80, 以其中的 40 个示例作为验证集,测试集大小为 400。实验结果如表 1 所示。

表 1 不同方法的实验结果比较

数据集		最佳个体网络	传统方法	本文方法
Friedman #3	误差均值	0.3176	0.3257	0.2252
	标准差	0.3217	0.1026	0.0805
Mexican Hat	误差均值	0.5676	0.6857	0.3609
	标准差	0.3217	0.0961	0.0374

表 1 中最佳个体网络指在构成神经网络集成的所有个体神经网络中,在测试集上误差最小的个体神经网络。从表 1 的实验结果可以看出:在两个数据集及运行实例下,本文提出的选择性神经网络集成方法都能取得比传统神经网络集成方法更小的泛化误差,稳定性也有所提高。

对于数据集 Friedman #3,用本节方法构造集成的过程中,平均每次选出 6 个网络组成集成。随着集成规模(即集成中个体网络的个数)的增大,集成的泛化误差的变化情况有待于进一步研究。在上述实验中,选择 5 个或 6 个个体网络组成集成,效果是比较好的。其原因可能是,如果集成规模太小,集成的泛化能力比单个网络没有明显的提高;如果集成规模太大,泛化能力相对较差的个体网络会被加入到集成中,从而会对集成结果产生负面的影响。

本文方法的个体选择算法在上述两个数据集上的运行时间曲线见图 1。图中 x 轴表示实验序号, y 轴表示运行时间。本文方法的个体选择算法在数据集 Friedman #3 上的平均运

行时间为 0.4821 秒,在 Mexican Hat 数据集上的平均运行时间为 0.5457 秒,其运行效率是比较高的。

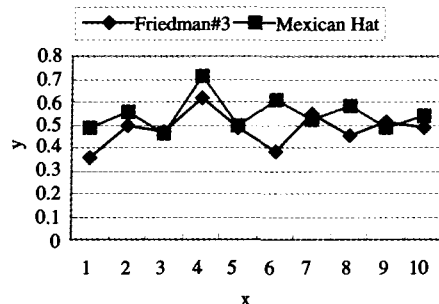


图1 个体选择算法运行时间情况

结论 本文提出一种基于约束规划的选择性神经网络集成方法。该方法一般情况下能够得到最优解,且设计过程简单,效率较高,能以较小的代价提高神经网络集成的泛化能力。但该方法也存在一定的问题,可能会遇到求相关系数矩阵的逆阵的问题,而实际应用中,各网络的误差可能有较大的

关联,这样个体网络的相关度矩阵是线性相关的,这时相关度矩阵的逆阵求解是相当困难的。为了避免这种情况,可以用一些方法^[6,7]训练个体网络,使它们彼此之间相关度较小。

参考文献

- 1 Krogh A, Vedelsby J. Neural network ensembles, cross validation, and active learning. In: Tesauro G, Touretzky D, Leen T, eds. *Advances in Neural Information Processing Systems*, Cambridge, MA: MIT Press, 1995, 7: 231~238
- 2 Partridge D, Yates W B. Engineering Multiversion Neural-net Systems. *Neural Computation*, 1996, 8: 869~893
- 3 Wu Jianxin, Zhou Zhihua, Chen Zhaoqian. Ensemble of GA-based Selective Neural Network Ensembles. In: *Proceedings of the 8th International Conference on Neural Information Processing (ICONP'01)*, Shanghai, China, 2001, 3: 1447~1482
- 4 周志华. 神经计算中若干问题的研究: [博士学位论文]. 南京大学, 2000
- 5 Friedman J. Multivariate adaptive regression splines. *Annals of Statistics*, 1991, 19(1): 1~141
- 6 Rosen B E. Ensemble learning using decorrelated neural networks. *Connection Science*, 1996, 8(3-4): 373~384
- 7 Turner K, Ghosh J. Error Correlation and Error Reduction in Ensemble Classifiers. *Connection Science*, Special issue on combining artificial neural networks: ensemble approaches, 1996, 8(3-4): 385~404

(上接第 138 页)

(1)对于 Stack-Tree-Desc 算法,无论是祖先/后裔关系还是双亲/孩子关系的结构连接,它的 CPU 时间和 I/O 复杂度均为 $O(|Alist| + |Dlist| + |Output|)$,对 Alist 和 tDlist 列表只需要分别扫描一次,连接结果立即被输出;栈的大小是 XML 文档树的高度。

(2)Stack-Tree-Anc 算法,无论是祖先/后裔关系还是双亲/孩子关系的结构连接,它的 CPU 时间和 I/O 复杂度均为 $O(|Alist| + |Dlist| + |Output|)$,对 Alist 和 Dlist 列表只需要分别扫描一次,连接结果需要进行缓存;栈的大小是 XML 文档树的高度。

模式匹配方法有两种:一是基于树遍历的模式匹配方法,二是基于集合的匹配方法。基于树遍历的匹配方法需要对 XML 文档树遍历,每一个 XML 文档树结点都需要被访问一次,因此可能出现许多没有必要的遍历扫描,以前的工作已经对模式树匹配问题做了大量研究。Twig 通过分解模式树为二元结构关系,分别连接,再把二元结果合并起来处理模式树匹配,但是会产生大量无用的中间结果。文[6]中提到的 TwigStack 算法通过维护多个栈并将它们链接起来,按流水线的方式实现小枝模式的所有结构连接。这样,可以将整棵模式树一起匹配处理,避免保存无用中间结果,它对于只有祖先后裔边的模式树匹配是最优的。但是它没有使用索引信息,必须对模式树中所有边都进行连接,而且它对于带有父子边的模式树不能保证最优。文[7]在 TwigStack 算法基础上使用了 XR-Tree 索引,可以跳过不参与连接的结点,但它也有 TwigStack 的缺点。

结论 目前,国内外科研人员在在这方面做了大量的研究工作。中国人民大学孟小峰等提出了基于区域划分的 XML 结构连接^[8],该算法基于任务分解的思想,利用区域编码的特点对输入集合进行划分。该算法在输入数据无序或没有索引的情况下优于现有的排序合并算法,可以为查询计划提供更多的选择。江西财经学院万常选等提出的基于区间编码的 XML 索引结构有效实现结构连接^[10],给出了一个 XML 树数据模型的形式化定义。文[9]将编码方案、逆序列表和路径索引的思想相结合,提出了一种改进的 XML 数据的索引结构,给出了两个实现双亲/孩子关系和拥有关系的结构连接算

法,它们最多只需要对参与连接的两个列表分别进行一次扫描,并且能够根据双亲结构信息等利用 B^+ 树索引尽可能多地跳过不需要参与连接的元素结点。通过国内外科研人员对 XML 数据库结构连接研究的不断深入,会有更多、更有效的连接算法出现,为 XML 数据库的查询优化奠定良好的基础。

参考文献

- 1 Zhang C, Naughton J, DeWitt D, et al. On Supporting Containment Queries in Relational Database Management Systems. In: Mehrotra S, et al. eds. *Proceedings of the 20th ACM SIGMOD International Conference on Management of Data*, Santa Barbara, California, USA. MAY, 2001. New York: ACM Press, 2001. 426~437
- 2 刘云生,万常选,徐升华. 基于关系数据库有效地实现 PRE 查询. *小型微型计算机系统*, 2003, 24(10): 1764~1771
- 3 Al-Khalifa S, Jagadish H V, Koudas N, et al. Structural Joins: A Primitive for Efficient XML Query Pattern Matching. In: Hiong Ngu A H, et al. eds. *Proceedings of the 18th IEEE ICDE International Conference on Data Engineering*, San Jose, California, USA. February 26-March 1, 2002. Los Alamitos: IEEE Computer Society, 2002. 141~152
- 4 Chen Shu-Yao, Vagena Z, Zhang Donghui. Efficient Structural Joins on Indexed XML Documents. In: *Proceedings of the 28th VLDB International Conference on Very Large Database*, Hong Kong, China. August 2002. San Francisco: Morgan Kaufmann Publishers, 2002. 263~274
- 5 Jiang Haifeng, Lu Hongjun, Wang Wei, et al. XR-Tree: Indexing XML Data for Efficient Structural Joins. In: Casati F, et al. eds. *Proceedings of the 19th IEEE ICDE International Conference on Data Engineering*, Bangalore, India. March 2003. Los Alamitos: IEEE Computer Society, 2003. 253~264
- 6 Bruno N, Koudas N, Srivastava D. Holistic Twig Joins: Optimal XML Pattern Matching. In: Franklin M J, et al. eds. *Proceedings of the 21th ACM SIGMOD International Conference on Management of Data*, Madison, Wisconsin, USA. June 2002. New York: ACM Press, 2002. 310~321
- 7 Jiang Haifeng, Wang Wei, Lu Hongjun, et al. Holistic Twig Joins on Indexed XML Documents. In: Heuer A, et al. eds. *Proceedings of the 29th VLDB International Conference on Very Large Database*, Berlin, Germany. Sept. 2003. San Francisco: Morgan Kaufmann Publishers, 2003. 273~284
- 8 王静,孟小峰,王珊. 基于区域划分的 XML 结构连接. *软件学报*, 2004, 15(5): 720~729
- 9 Wang Wei, Jiang Haifeng, Lu Hongjun, et al. PbiTree Coding and Efficient Processing of Containment Joins. In: Casati F, et al. eds. *Proceeding of the 19th IEEE ICDE International Conference on Data Engineering*, Bangalore, India. March, 2003. Los Alamitos: IEEE Computer Society, 2003. 391~402
- 10 万常选,刘云生,徐升华,等. 基于区间编码的 XML 索引结构有效实现结构连接. *计算机学报*, 2005, 28(1)