

文章编号: 1004-4353(2020)02-0145-05

基于势函数聚类的改进 RBF 网络算法研究

黄朝辉, 闻辉*, 车艳

(莆田学院 信息工程学院, 福建 莆田 351100)

摘要: 为优化径向基函数(RBF)网络结构并改善网络泛化性能,提出了一种基于势函数聚类的改进 RBF 网络算法.首先使用势函数统计每个模式类别中的样本势值,以此实现样本空间中不同样本的势值度量;其次以增量学习的方式逐次完成对样本空间的有效覆盖,以此实现网络隐节点个数及参数的自动有效估计.最后将本文算法与 KMRBF、FCRBF、MRAN 以及 GAP-RBF 学习算法进行了实验对比,结果表明本文算法的网络分类精度更高,克服了 KMRBF 和 FCRBF 算法需人工调整网络隐节点来提高分类精度的问题,且比 GAP-RBF 和 MRAN 算法的网络结构更加简单.

关键词: 径向基函数; 增量学习; 势函数; 神经网络

中图分类号: TP183

文献标识码: A

Research on improved RBF network algorithm based on potential function clustering

HUANG Chaohui, WEN Hui*, CHE Yan

(Department of Information Engineering, Putian University, Putian 351100, China)

Abstract: To optimize the structure of radial basis function (RBF) network and improve its generalization performance, an improved RBF network based on potential function clustering is proposed. Firstly, the potential function is used to count the potential values of samples in each pattern category, so as to measure the potential values of different samples in the sample space. Secondly, the incremental learning method is used to cover the sample space step by step, so as to complete the automatic and effective estimation of the number of hidden nodes and parameters in the network. Finally, the presented algorithm is compared with KMRBF, FCRBF, MRAN and GAP-RBF learning algorithm by experiments. The results and experiments show that the classification accuracy of the algorithm is higher than that of KMRBF, FCRBF, GAP-RBF and MRAN, which overcomes the problem that KMRBF and FCRBF algorithm need many experiments to adjust the hidden nodes manually to obtain higher classification accuracy, and is more simple than the GAP-RBF and MRAN network structure.

Keywords: radial basis function; incremental learning; potential function; neural network

0 引言

径向基函数(RBF)网络因其结构简单、稳定性强的优点,目前被广泛应用于工业诊断、图像分

类以及语音识别等领域中.传统的优化 RBF 网络的算法包括 k 均值聚类(KMRBF)^[1]、模糊 C 均值聚类(FCRBF)^[2]和正交前向选择^[3],这些算法需要预先固定 RBF 网络隐节点的个数;但是固定

收稿日期: 2019-12-25

* 通信作者: 闻辉(1981—),男,博士,副教授,研究方向为机器学习、神经网络.

基金项目: 福建省自然科学基金资助项目(2019J01815);福建省教育厅中青年骨干教师教育科研项目(JT180486);莆田市科技局项目(2018RP4004,2018ZP10)

隐节点的个数往往使得网络结构与实际问题不匹配,致使网络泛化性能下降.为了改善网络结构与实际问题不匹配的问题,一些基于序学习的 RBF 网络结构优化算法被相继提出.例如:文献[4]提出了一种最小资源分配网络(MRAN)模型,该模型通过统计邻近样本之间的相关性影响来删除部分冗余的隐节点,但该方法未能考虑样本空间的整体分布.文献[5]提出了一种基于序列增加和裁剪的 RBF 网络(GAP-RBF)模型,该模型通过度量 RBF 隐节点的重要性来规划学习策略,但该算法需设定样本服从统一分布.文献[6]在 GAP-RBF 的基础上提出了一种泛化的 GAP-RBF (GGAP-RBF)模型,该模型虽然克服了 GAP-RBF 中样本需服从统一分布的问题,但是网络训练阶段的参数选取过于复杂.文献[7-8]提出了一种基于敏感度分析的 RBF 网络模型,该模型通过分析隐节点对整个网络的影响来动态调整 RBF 网络结构.上述文献研究的模型虽然可以建立自适应的 RBF 网络结构,但未能从整体上考虑不同样本之间的关联信息,因此网络的泛化性能会受到不同程度的影响.为此,本文提出一种基于势函数聚类的改进 RBF 网络算法,并通过双月人工数据集^[9]以及 UCI 机器学习数据库^[10]中的 Blood Transfusion、Diabetes 与 Image Segmentation 3 个基准数据集验证本文算法的有效性.

1 RBF 网络算法原理

RBF 网络由输入层、隐藏层和输出层 3 个模式层组成,其中输入层和隐藏层各节点的连接权值为 1.网络隐节点通常使用高斯核函数来实现对输入向量的映射.对于输入向量 $\mathbf{x} \in \mathbf{R}^h$,其经过隐节点的核映射后的输出可表示为

$$\varphi_k(\mathbf{x}) = \exp\left(-\frac{1}{2\sigma_k^2}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right), \quad k=1,2,\dots,K. \quad (1)$$

其中 $\boldsymbol{\mu}_k$ ($\boldsymbol{\mu}_k \in \mathbf{R}^h$) 和 σ_k 分别为网络第 k 个隐节点的中心和核宽, K 为 RBF 网络隐节点的个数.

RBF 网络的输出可表示为

$$F(\mathbf{x}_k) = \sum_{k=1}^K \omega_k \varphi_k(\mathbf{x}), \quad (2)$$

其中 ω_k 为第 k 个隐节点与输出层对应节点之间的连接权值.

2 基于势函数聚类的改进 RBF 网络算法的实现

由于势函数可以度量样本空间中的两个不同向量随距离的变化情况,因此本文提出基于势函数聚类的改进 RBF 网络学习算法,即通过对样本空间的学习来增量生成不同的隐节点,以此实现对样本空间不同区域的有效覆盖.设 \mathbf{x}_1 和 \mathbf{x}_2 为样本空间的两个向量, $\gamma(\mathbf{x}_1, \mathbf{x}_2)$ 为由这两个样本建立的势函数. $\gamma(\mathbf{x}_1, \mathbf{x}_2)$ 的表达式为

$$\gamma(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{1 + T \cdot d^2(\mathbf{x}_1, \mathbf{x}_2)}. \quad (3)$$

其中: T 为常数,用以控制势函数随距离变化的衰减速度; $d(\mathbf{x}_1, \mathbf{x}_2)$ 为 $\mathbf{x}_1, \mathbf{x}_2$ 之间的距离.

给定训练样本集合 D , 其中 \mathbf{x} 为 D 中的任意一个向量, h_i 为与 \mathbf{x} 相关的类别标签, $h_i \in R^c$, c 为模式类别的个数.设 D_i 为 D 中标签为 h_i 的样本集合, $D_i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, \dots, \mathbf{x}_{N_i}^i\}$, 其中 N_i 为 D_i 中的样本个数, 则有 $D = \bigcup_{i=1}^c D_i$, 这里 $D_i \cap D_j = \emptyset, \forall i \neq j$. 对 D_i 中的任意一对样本向量 \mathbf{x}_p^i 和 \mathbf{x}_q^i , 其作用势为

$$\gamma(\mathbf{x}_p^i, \mathbf{x}_q^i) = \frac{1}{1 + T \cdot d^2(\mathbf{x}_p^i, \mathbf{x}_q^i)}, \quad (4)$$

式中 $p, q = 1, 2, \dots, N_i$ 且 $p \neq q$. 若以 \mathbf{x}_q^i 作为基准样本, 并累加 D_i 中其他各样本向量对 \mathbf{x}_q^i 的作用势可得

$$\kappa_q^i = \sum_{p=1, p \neq q}^{N_i} \gamma(\mathbf{x}_p^i, \mathbf{x}_q^i). \quad (5)$$

其中 κ_q^i 表示样本 \mathbf{x}_q^i 在第 i 个模式类别中的势值, 它反映的是该样本所在区域的稀疏程度. 当 κ_q^i 较小时, 可将 \mathbf{x}_q^i 所在的区域认定为相对稀疏, 反之可将 \mathbf{x}_q^i 所在的区域认定为相对稠密. 由公式(5)可得 D_i 中各样本的势值集合为 $\kappa^i = \{\kappa_1^i, \kappa_2^i, \dots, \kappa_{N_i}^i\}$.

在 RBF 网络中, 隐节点使用高斯核函数将样本空间划分为不同的子空间, 其所划分的子空间的个数即为 RBF 隐节点的个数. 为构建不同参数的高斯核函数以实现样本空间的逐次覆盖, 需在每次统计样本势值后, 将 κ^i 中势值最大的对应样本作为高斯核中心. 假设 $\kappa_u^i = \max\{\kappa_1^i, \kappa_2^i, \dots, \kappa_{N_i}^i\}$, 则可将样本 \mathbf{x}_u^i 作为所构建的高斯核中心, 即有

$$\mu_k = x_u^i, \quad (6)$$

式中 k 为在 RBF 网络隐藏层中生成的高斯核个数. 将高斯核宽 σ 与 μ_k 相结合, 即可建立一个对应的高斯核用以实现对样本空间某个局部区域的有效覆盖.

更新高斯核时需消除当前高斯核所覆盖的局部区域的样本势值. 对 D_i 中任意一个样本 x_q^i , 其更新过程可表示为

$$\kappa_q^{i'} = \kappa_q^i - \kappa_u^i \cdot \exp\left(-\frac{1}{2\sigma_k^2} \|x_q^i - x_u^i\|^2\right), \quad (7)$$

$$q = 1, 2, \dots, N_i, q \neq u.$$

其中: $\kappa_q^{i'}$ 为更新后的样本势值, σ_k 为高斯核宽, κ_u^i 为第 i 个模式类别中更新前的最大样本势值, $\kappa_u^i \cdot \exp\left(-\frac{1}{2\sigma_k^2} \|x_q^i - x_u^i\|^2\right)$ 表示当前的高斯核中心 x_u^i 对样本 x_q^i 势值的影响. 由高斯核函数的衰减特性可知, 当 D_i 中的样本 x_q^i 距离高斯核中心 x_u^i 较远时, x_u^i 对样本 x_q^i 的势值影响较小, 即 x_q^i 的势值被抵消得较小; 反之, x_q^i 的势值被抵消得较大. 由此知, D_i 中的各样本势值可被有效更新. 当 D_i 中更新后的样本势值满足 $\max\{\kappa_1^{i'}, \kappa_2^{i'}, \dots, \kappa_{N_i}^{i'}\} > \delta$ (δ 为阈值) 时, 本文算法即可从更新后的势值集合中选取势值最大的对应样本作为下一个高斯核的中心, 且可使下一个高斯核中心与当前的高斯核中心 x_u^i 的距离较远, 由此实现对样本空间不同区域的逐次 (从相对稠密区域到相对稀疏区域) 覆盖. 完成覆盖后, 算法对下一个样本集合 D_j ($j \neq i$) 进行学习, 直到学完所有训练样本, 完成整个样本空间的有效覆盖后, 即可实现对 RBF 网络隐节点个数及对应参数的有效估计. 图 1 为本文采用势函数聚类算法构建的 RBF 网络隐节点的示意图. 图 1 中的圆圈表示本文算法对样本空间不同局部区域的覆盖, 且每覆盖一次就自动增量生成一个相对应的 RBF 隐节点.

图 2 为本文提出的基于势函数聚类的改进 RBF 网络隐节点学习算法, 算法中对输出权值 $\{\omega_k\}_{k=1}^K$ 的优化采用 RLS^[1] 算法来实现.

3 实验及分析

为了验证本文所提算法的性能, 将本文所提算法与 4 种典型的优化 RBF 网络算法 (KMRBF、FCRBF、MRAN、GAP-RBF) 进行实验对比. 表 1

为不同分类数据集的信息. 图 3 为双月人工数据集的结构图, 图 3 中 $r=10$, $\omega=6$, $d=-6$. 实验中各数据集中的样本全部归一化到 $[-1, 1]$ 范围内. 将实验参数 T 和 δ 设置为 $T=1$, $\delta=0.01$, σ 的取值为 $0.1 \sim 0.6$. 实验运行环境为 Intel(R) Core (TM) i5-7500, 3.4 GHZ CPU, 8 G RAM, MATLAB 2013.

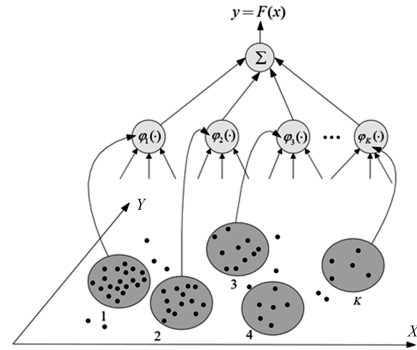


图 1 基于势函数聚类增量构建的 RBF 网络隐节点示意图

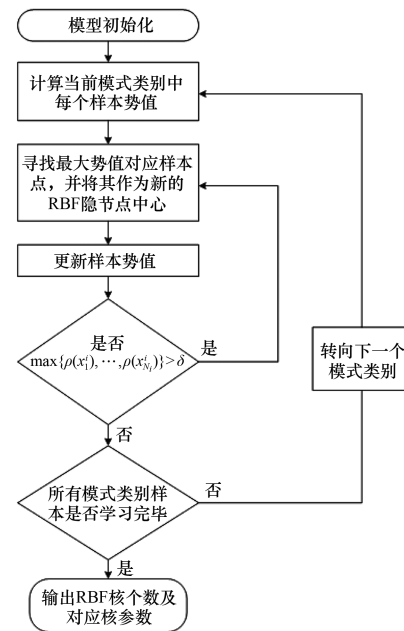


图 2 基于势函数聚类的改进 RBF 网络隐节点学习算法

3.1 双月数据集下本文算法的学习效果

图 4 为训练集样本数为 500 且核宽参数取不同值时的本文算法的学习效果. 图 4 中每一个圆圈表示本文所建立的高斯核对样本空间的一次覆盖, 且每一个圆圈都生成一个 RBF 网络隐节点. 由图 4 可以看出, 样本空间的覆盖区域随核宽参数的增加而增大, 所生成的网络隐节点的个数及中心参数也随之改变, 这表明本文所提算法具有

良好的自适应性,可应用于复杂数据分类下的 RBF 网络隐节点个数以及参数估计.

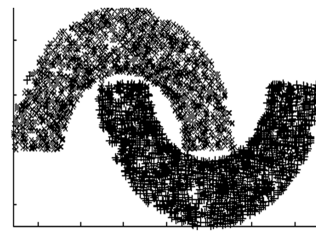
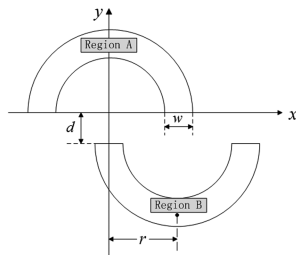
3.2 双月数据集下不同算法的性能对比

图 5 为双月数据集下不同算法的性能.从图 5(a)可以看出,随着训练样本个数的增加,本文算法生成的隐节点个数明显少于 MRAN 和 GAP-RBF 算法,但多于 KMRBF 算法.虽然本文算法生成的隐节点个数多于 KMRBF 算法,但因本文算法的隐节点个数是自动生成的,而 KMRBF 算

法的隐节点个数需经过多次实验和手动设定才能实现,因此本文算法对样本空间具有更好地适应性.需要说明的是,在本实验中当 FCRBF 算法与 KMRBF 算法的隐节点个数相同时,FCRBF 算法可以获取相对更高的分类精度,因此在图 5(a)中未给出 FCRBF 算法.从图 5(b)中可以看出,本文算法的分类精度优于 KMRBF、FCRBF、GAP-RBF 和 MRAN 算法,尤其是在训练样本个数为 200~500 时,本文算法的分类精度相对更高.

表 1 不同分类数据集的信息

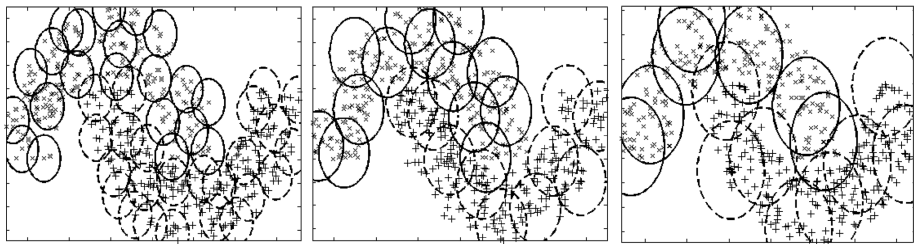
数据集	类别	特征维数	训练集样本数	验证集样本数	测试集样本数
双月	2	2	200~2 000	500	4 000
BloodTransfusion	2	4	374	187	187
Diabetes	2	8	576	100	92
Image Segmentation	7	19	210	1 000	1 100



(a)分类原型

(b)分类数据集

图 3 双月人工数据集结构图

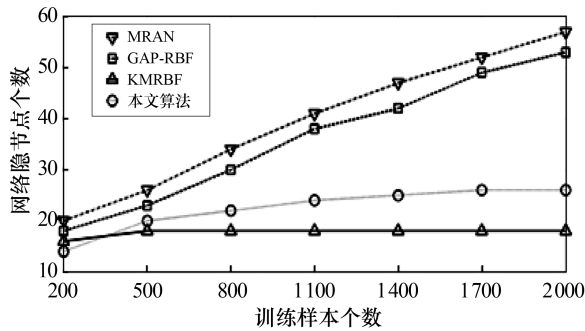


(a) $\sigma=2$

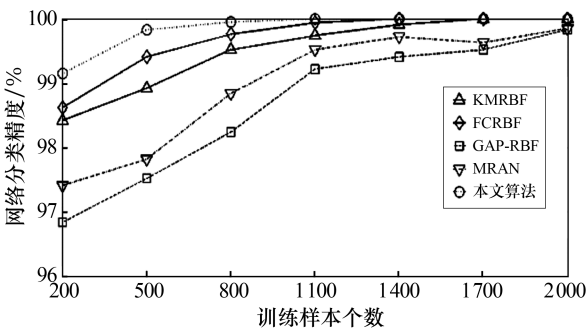
(b) $\sigma=3$

(c) $\sigma=4$

图 4 取不同核宽参数值时本文算法对训练样本空间的学习效果



(a)不同训练样本数量下的 RBF 网络的隐节点数量



(b)不同训练样本数量下的 RBF 网络的分类精度

图 5 双月数据集下不同算法的性能

3.3 基准数据集下不同算法的性能对比

在 UCI 基准数据集下,本文算法与其他 4 种算法的性能对比见表 2—表 4。由表 2—表 4 可以看出:本文算法的测试精度均优于其他 4 种算法。在隐节点个数方面,本文算法优于 GAP-RBF 算法和 MRAN 算法,但低于 KMRBF 算法和 FCRBF 算法。在 Blood Transfusion 基准数据集下,本文算法的训练时间优于 GAP-RBF 和 MRAN 算法,但略低于 KMRBF 和 FCRBF 算法;在 Diabetes 和 Image Segmentation 基准数据集下,本文算法的训练时间均优于其他 4 种算法。

表 2 Blood Transfusion 基准数据集下不同算法的性能

算法	训练时间/s	测试精度/%	隐节点个数
KMRBF	4.28	75.74	36
FCRBF	4.45	76.12	32
GAP-RBF	12.32	73.21	177
MRAN	10.59	73.07	143
本文算法	4.81	76.63	116

表 3 Diabetes 基准数据集下不同算法的性能

算法	训练时间/s	测试精度/%	隐节点个数
KMRBF	2.47	76.25	20
FCRBF	3.42	76.96	20
GAP-RBF	2.93	75.39	41
MRAN	2.21	76.63	48
本文算法	1.32	78.56	35

表 4 Image Segmentation 基准数据集下不同算法的性能

算法	训练时间/s	测试精度/%	隐节点个数
KMRBF	5.29	73.75	52
FCRBF	6.32	74.52	46
GAP-RBF	7.73	65.45	137
MRAN	8.75	68.62	116
本文算法	2.38	90.68	77

4 结论

研究表明,本文所提出的基于势函数聚类的改进 RBF 网络算法可有效利用样本空间的分布信息,并据此自动确定 RBF 隐节点的个数及参数。本文算法的分类精度优于 KMRBF、FCRBF、GAP-RBF、MRAN 4 种算法。与 KMRBF 和 FCRBF 算法相比,本文算法可以有效克服 RBF 网络结构

需手动调整的问题;在 Diabetes 和 Image Segmentation 基准数据集上,本文算法的训练时间均优于上述 4 种算法;与 GAP-RBF 和 MRAN 算法相比,本文算法所得的网络结构更加简单,训练时间也相对更优。本文研究结果为优化 RBF 网络隐节点核参数提供了新的思路。本文在研究中未能考虑核宽的动态调整情况,因此在今后的研究中我们将引入核宽覆盖因子,建立样本空间的局部优化覆盖,以完善本文模型。

参考文献:

- [1] MOODY J, DARKEN C J. Fast learning in networks of locally-tuned processing[J]. Neural Computation, 1989,1(2):281-294.
- [2] 张爱科,符保龙,李辉. 基于改进的模糊聚类 RBF 网络集成的文本分类方法[J]. 四川大学学报(自然科学版),2012,49(6):1235-1239.
- [3] RAITOHARJU J, KIRANYAZ S, GABBOUJ M. Training radial basis function neural networks for classification via class-specific clustering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016,27(12):2458-2471.
- [4] LU Y W, SUNDARARAJAN N, SARATCHANDRAN P. A sequential learning scheme for function approximation using minimal radial basis function[J]. Neural Computation, 1997,9(2):461-478.
- [5] HUANG G B, SARATCHANDRAN P, SUNDARARAJAN N. An efficient sequential learning algorithm for growing and pruning RBF (GAP-RBF) networks[J]. IEEE Transactions on Systems, Man and Cybernetics. Part B: Cybernetics, 2004,34(6):2284-2292.
- [6] HUANG G B, SARATCHANDRAN P, SUNDARARAJAN N. A generalized growing and pruning RBF (GAP-RBF) neural network for function approximation[J]. IEEE Transactions on Neural Networks, 2005,16(1):57-67.
- [7] 乔俊飞,韩红桂. RBF 神经网络的结构动态优化设计[J]. 自动化学报,2010,36(6):865-872.
- [8] 韩红桂,乔俊飞,薄迎春. 基于信息强度的 RBF 神经网络结构设计研究[J]. 自动化学报,2012,38(7):1083-1090.
- [9] BLAKE C, MERZ C. UCI repository of machine learning databases[EB/OL]. [2019-06-03]. <http://archive.ics.uci.edu/ml/35>.
- [10] HAYKIN S. Neural Networks and Learning Machines[M]. Third Edition. Beijing: China Machine Press, 2009:61-63.