优化的 RBF 网络在特征选择中的应用研究*

朱颢东[1,2],钟 勇[1,2]

(1.中国科学院成都计算机应用研究所,四川 成都 610041;

(2.中国科学院研究生院,北京 100039)

摘 要:提出了一个自适应量子粒子群优化算法,用于训练 RBF 网络的基函数中心和宽度,并结合最小二乘法计算网络权值,对 RBF 网络的泛化能力进行改进并用于特征选择。实验结果表明,采用自适应量子粒子群优化算法获得的 RBF 网络模型不但具有很强的泛化能力,而且具有良好的稳定性,能够选择出较优秀的特征子集。

关键词:特征选择;文本分类;RBF 神经网络;量子粒子群优化;最小二乘法;

中图分类号:TP301 **文献标识码**:A

Study of optimized RBF network in feature selection

ZHU Hao-Dong[1,2], ZHONG Yong[1,2]

(1.Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China;

(2.The Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)

Abstract: An adaptive quantum-behaved particle swarm optimization (AQPSO) algorithm is firstly proposed in order to improve the performance of RBF network. By applying AQPSO algorithm to train the central position and width of the basis function adopted in the RBF network, and computing the weights of the network with least-square method, the generalization ability of the RBF neural network is improved, and then applied to select features. Experimental results show that obtained network model not only has good generalization properties, but also has better stability. It illustrates that RBF neural network with AQPSO algorithm can acquire the feature subset which is more representative.

Key words: feature selection; text categorization; RBF neural network; quantum-behaved particle swarm optimization; least-square method

在文本分类中,文本通常是以向量形式来表示的,其特点是高维性和稀疏性[1]。而在中文文本分类中,通常采用词条作为最小的独立语义载体,原始的特征空间可能由出现在文章中的全部词条构成。由于中文的词条总数有 20 多万条,这使得其高维性和稀疏性更加明显,这样就大大限制了分类算法的选择空间,降低了分类算法的效率和精度。为此,寻找一种高效的特征选择方法,以降低特征空间维数、避免维数灾难,提高文本分类的效率和精度,成为文本自动分类中亟待解决的重要问题[2-4]。

本文首先提出了一个自适应量子粒子群优化算法,用于训练 RBF 网络的基函数中心和宽度,并结合最小二乘法计算网络权值,对 RBF 网络的泛化能力进行改进。然后把该 RBF 神经网络用于特征选择。实验结果表明,采用自适应量子粒子群优化算法获得的 RBF 网络模型不但具有很强的泛化能力,而且具有良好的稳定性,能够选择出较优秀的特征子集。

1 RBF 神经网络简介

RBF 网络是一种 3 层前馈神经网络[5-7]。由输入层、隐层和输出层组成。网络输出层的输出计算公式为:

$$y_i=\sum_{k=1}^{N} w_{ik}\varphi_k(\boldsymbol{x},c_k)=\sum_{k=1}^{N} w_{ik}\varphi_k(\|\boldsymbol{x}-c_k\|_2) \quad (1)$$

这里 $\boldsymbol{x}\in R^{n\times l}$ 是输入矢量;$\varphi_k(\cdot)$ 是从正实数集合到实数集合的函数,此函数的给出形式较多,这里采用高斯函数:

* 基金项目:四川省科技计划项目(2008GZ0003)

$$\varphi_k(\boldsymbol{x})=\exp(-\boldsymbol{x}^2/\sigma) \quad (2)$$

$\|\cdot\|_2$ 表示欧几里德范式；w_{ik} 是输出层的权值；N 表示隐含层神经元的个数；$c_k \in R^{n\times l}$ 表示输入向量空间的 RBF 网络的中心(简称中心)。对于隐含层的每一个神经元，要计算与其关联的中心和网络输入间的欧几里德距离。隐含层神经元的输出是该距离的非线性函数。隐含层输出的加权和是 RBF 网络的输出[5-7]。

由公式(1)和(2)可见，如果知道了 RBF 网络的隐层节点数 N、中心 c_k 和 σ 宽度，RBF 网络从输入到输出就成了一个线性方程组，此时连接权值的学习可采用最小二乘法求解。因此，只要确定了 N、c_k 和 σ，RBF 网络模型也建立好了。而对 RBF 网络的隐层节点个数，本文采用了 SOM(Self Organization Mapping)网络的聚类算法来确定。对中心 c_k 和宽度 σ 则由本文提出的自适应量子粒子群优化算法来确定。

2 优化的粒子群算法

2.1 量子粒子群优化算法(QPSO)

粒子群优化(PSO)算法是一种基于群体智能的随机搜索算法，其粒子在经典力学空间中飞行，由于飞行轨迹确定，因此搜索空间有限，易陷入局部极值[8]。为了避免这个缺陷参考文献[9]从量子力学的角度出发提出具有量子行为的粒子群优化算法(QPSO)。由于在量子空间中的粒子移动时没有确定的轨迹，使粒子可以在整个可能解空间中随机搜索全局最优解，因此 QPSO 算法的全局搜索能力远远优于 PSO 算法[9]。

由于在量子空间中，粒子的位置和速度不能同时确定，因此可以通过波函数(波函数的平方是粒子在空间中某一点出现的概率密度)来描述粒子的状态，并通过求解薛定谔方程得到粒子在空间某一点出现的概率密度函数，随后通过蒙特卡罗随机模拟的方式得到量子空间中粒子的位置方程，如下式所示[10]：

$$p=a\times p_b(i)+(1-a)\times g_b \quad (3)$$

$$m_b=\frac{1}{N}\sum_{i=1}^{N}p_b(i) \quad (4)$$

$$b=1-i_t/i_{t\ \max}\times 0.5 \quad (5)$$

$$p_{pos}=p\pm b\times|m_b-p_{pos}|\times\ln\frac{1}{u} \quad (6)$$

其中 p 为 p_b 与 g_b 之间的随机位置；m_b 是所有粒子个体最佳位置 p_b 的平均值；N 为粒子个数；b 为收缩扩张系数，在 QPSO 算法收敛的过程中线性减小；i_t 为当前迭代次数；$i_{t\ \max}$ 为设定的最大迭代次数；p_{pos} 是粒子的当前位置；a 和 u 都是 0~1 之间的随机数，当 $u>=0.5$ 时，公式(5)取“—”号，当 $0<u<0.5$ 时，公式(5)取“+”号。

2.2 自适应量子粒子群优化算法(AQPSO)

在 QPSO 算法中，当 p_b 和 g_b 很接近时意味着粒子的参数 P 很小，于是粒子的搜索范围也变得很小。这样，粒子群的进化就会停滞；如果这个时候粒子群的当前最佳位置处于一个局部最优解，那么整个粒子群就会趋于早熟收敛[10]。

在 QPSO 算法中，只有一个收缩扩张系数 b，对这个参数的选择和控制是非常重要的，它关系到整个算法的收敛性能[10]。在参考文献[10]中已经证明，当 $b<1.7$ 时，粒子收敛，靠近粒子群的当前最佳位置；当 $b>1.8$ 时，粒子发散，远离粒子群的当前最佳位置。从公式(5)可以看出收缩扩张系数 b 在粒子进化过程随着进化代数的增加而线性减小，这种固定的变化并不能自适应避免早熟趋势。对此，本文做出如下改进：

$$h=f_{bfit2}/f_{fitness}(i) \quad \text{若 } h<0.5 \quad b=2\times h \quad (7)$$

$$\text{若 } h\geqslant 0.5 \quad b=1+h \quad (8)$$

其中 f_{bfit2} 为上一代群体获得最佳位置 g_b 的适应度；$f_{fitness}(i)$ 为第 i 个粒子的当前适应度。h 为两者的比值，h 越小，说明粒子越远离粒子群的当前最佳位置；h 越大，说明粒子越靠近粒子群的当前最佳位置。本文以 h 的值是否小于 0.5 为分界，如果 $h<0.5$，说明粒子远离群体最佳位置 g_b，收缩扩张系数 b 应该小于 1.7，使它收敛。因此将 b 值设为 $2\times h$，使它不超过 1；$h\geqslant 0.5$，说明粒子靠近群体的当前最佳位置 g_b，因此将 b 值设为 $1+h$，增加其大于 1.8 的概率，使它尽量发散，扩大搜索范围。

3 AQPSO 算法对 RBF 网络的优化

用 AQPSO 算法训练 RBF 神经网络时，首先要用向量形式表示 RBF 网络的学习过程中需要调整的 2 个训练参数：(1)径向基函数的数据中心 c。(2)径向基函数的扩展常数，即宽度 σ。假设采用 SOM 聚类算法得到 RBF 网络有个隐层节点，粒子群的规模，即粒子的个数为 N，则对粒子参数编码格式如表 1 所示，粒子群编码格式如表 2 所示。粒子参数维数 $D=2m$，每个粒子用 1 个 $2m$ 维的向量来表示对应的 m 个径向基函数的数据中心和宽度，则粒子在 $N\times D$ 维的解空间 POP 中搜索群体的最佳位置，粒子群体的最佳位置对应 RBF 网络中的最优的数据中心值和宽度。

表 1 粒子编码格式

$c_1, c_2, \cdots c_m$	$\delta_1, \delta_2, \cdots \delta_m$

表 2 粒子群编码格式

$c_{11}, c_{12}, \cdots c_{1m}$	$\delta_{11}, \delta_{12}, \cdots \delta_{1m}$
$c_{21}, c_{22}, \cdots c_{2m}$	$\delta_{21}, \delta_{22}, \cdots \delta_{2m}$
…	…
$c_{N1}, c_{N2}, \cdots c_{Nm}$	$\delta_{N1}, \delta_{N2}, \cdots \delta_{Nm}$

计算粒子群体的最佳位置需要比较粒子的适应度，本文以每个粒子对应的网络参数在训练集上产生的均方差 E_{MS} 作为粒子的适应度的目标函数，则适应度越小，E_{MS} 就越小，网络对数据的拟合程度就越高。粒子的适应度 $f_{fitness}$ 可用下面的公式计算：

$$f_{fitness}=\frac{1}{2N}\sum_{i=1}^{N}(y_i-t_i)^2 \quad (9)$$

其中 y_i 为第 i 个粒子的实际输出值；t_i 为第 i 个粒子的期望输出值。一旦粒子搜索完成，找到的粒子群中适应度最小者，即拥有最佳位置 g_b，则对应的隐层节点的最优的数据中心和宽度也就确定了。对于 RBF 网络隐

层到输出层网络连接权值向量则可以使用最小二乘法(LMS)直接计算得到。

具体的优化 RBF 网络模型实现如下:

(1)初始化粒子群体 POP、每个粒子的最佳位置 $p_b(0)=\varnothing$、粒子群最佳位置 $g_b=\varnothing$、粒子的适应度 $f_{fitness}(0)=0$、当前粒子群的最佳适应度 $f_{bfit1}=0$、上一代粒子群的最佳适应度 $f_{bfit2}=0$ 和预设精度 $\varepsilon=0.09$、最大迭代次数 $i_{t\ max}=200$,$i=1$。

(2)根据当前粒子 i 的位置(得到网络的中心和宽度),结合最小二乘法(得到网络的连接权值)计算出粒子 i 对所有训练样本的适应度;并比较粒子 i 的适应 $f_{fitness}(i)$ 和整个粒子群体的适应度 f_{bfit1},如果 $f_{fitness}(i)<f_{bfit1}$,则更新粒子 i 最佳位置 $p_b(i)$。

(3)判断所有粒子是否完成搜索,是则转(4),否则返回(2)。

(4)比较当前群体的最佳适应度 f_{bfit1} 和上一代群体的最佳适应度 f_{bfit2},若 $f_{bfit1}<f_{bfit2}$,则更新粒子群最优位置 g_b 和粒子群的最佳适应度 f_{bfit1}。

(5)判断粒子群中最佳的适应度即最小 E_{MS},是否小于预设精度 ε,是则转(8),否则转(6)。

(6)$i=i+1$,如果 $i\geqslant i_{t\ max}$,则转(8),否则返回(7)。

(7)根据公式(8)至(9)更新每个粒子的位置,生成新的粒子群,返回(2)。

(8)RBF 网络训练完成,输出粒子群最佳位置 g_b。其中,$g_b(1:m)$ 对应 RBF 网络最优的 m 个数据中心,$g_b(m+1:2\times m)$ 对应 RBF 网络最优的 m 个宽度。用 LMS 计算出网络连接权值,建立基于 AQPSO 算法的 RBF 网络预测模型。

4 本文特征选择方法

本文特征选择方法使用了本文提出的基于 AQPSO 算法优化的 RBF 神经网络。简单过程如下:

训练样本首先经过分词、特征提取得到原始特征集;然后利用参考文献[2]提出的优化的文档频方法先过滤掉一些词条(最小词频数阈值 $n=2$,最小文档数阈值 $m=5$,这时的特征集为初选特征集),来降低特征空间维数,从而降低 RBF 网络的输入层的单元数,以减少该网络的训练耗时。最后用本文所给的优化的 RBF 网络进行特征优选,从而选择出较优的特征子集。

使用优化的 RBF 网络进行特征优选的方法如下:把每个训练样本表示成向量的形式,每个初选特征(经过优化的文档频方法筛选的特征)在这个向量中对应一个权值。本文取该特征在这个文本中出现的次数和这个文本所属类的总训练文本中包含该特征的文本数的乘积为权值。将所有文本向量(相当于初始粒子群)作为训练样本,RBF 网络的输入层神经元个数等于初选特征数;输出层神经元个数等于训练文本的类别个数;隐含层神经元个数相对固定(以网络的泛化性和训练效率确定)。经过训练后,存在一些较大权值对应的隐含层神经元,与其相连接的输入层神经元所代表的特征即为特征,它们的并集就是优选的特征子集。

5 实验例证

5.1 实验语料库

在中文文本分类方面,经过分析和比较,本文选用的分类语料库是复旦大学中文文本分类语料库。该语料库由复旦大学计算机信息与技术系国际数据库中心自然语言处理小组构建,语料文档全部采自互联网,可以免费下载,网址为:http://www.nlp.org.cn/categories/default.php?cat_id=16。

复旦大学中文文本分类语料库中包含 20 个类别,分为训练文档集和测试文档集 2 个部分。每个部分都包括 20 个的子目录,相同类别的文档存放在一个对应的子目录下;每个存储文件只包含 1 篇文档,所有文档均以文件名作为唯一编号。共有 19 637 篇文档,其中训练文档 9 804 篇,测试文档 9 833 篇;训练文档和测试文档基本按照 1:1 的比例来划分。去除部分重复文档和损坏文档后,共保留文档 14 378 篇,其中训练文档 8 214 篇,测试文档 6 164 篇,跨类别的重复文档没有考虑,即 1 篇文档只属于 1 个类别。该语料库中的文档的类别分布情况是不均匀的。其中,训练文档最多的类 Economy 有 1 369 篇训练文档,而训练文档最少的类 Communication 有 25 篇训练文档;同时,训练文档数少于 100 篇的稀有类别共有 11 个。训练文档集和测试文档集之间互不重叠。本文只取前 10 个类的部分文档,其类别文档分布如表 3 所示。

表 3 文档分布

类别	训练文档数目	测试文档数目
经济	480	419
体育	584	489
计算机	628	591
政治	573	482
农业	547	435
环境	405	371
艺术	510	286
太空	506	248
历史	466	468
军事	74	75

5.2 实验环境及参数设置

实验设备是 1 台普通计算机:操作系统为 Microsoft Windows XP Professional (SP2),CPU 规格为 Intel (R) Celeron(R) CPU 2.40 GHz,内存 512 MB,硬盘 80 GB。

进行中文分词处理时,采用的是中科院计算所开源项目“汉语词法分析系统 ICTCLAS”系统。

实验使用的软件工具是 Weka,这是纽西兰的 Waikato 大学开发的数据挖掘相关的一系列机器学习算法。实现语言是 Java,可以直接调用,也可以在代码中调用。Weka 包括数据预处理、分类、回归分析、聚类、关联

规则、可视化等工具，对机器学习和数据挖掘的研究工作很有帮助，是开源项目，网址为：http://www.cs.waikato.ac.nz/ml/weka/。实验使用的计算工具为 MATLAB 7.0。

5.3 实验所用分类器及其评价标准

本实验旨在比较本文方法与信息增益(IG)、x^2 统计量(CHI)、互信息(MI)等 3 种特征选择方法对后续文本分类精度的影响，因此本实验在各种特征选择方法后采用相同的分类器对文本进行分类。本实验中使用 KNN 分类器来比较这几种特征选择方法(K 设置为 10)。

为了评价实验效果，实验中选择分类正确率和召回率作为评价标准：准确率(Precision)=$a/(a+b)$，它是所判断的文本与人工分类文本吻合的文本所占的比率；召回率(Recall)=$a/(a+c)$，是人工分类结果应有的文本与分类系统吻合的文本所占的比率。在实际中，查准率比查全率重要。其中 a、b、c 代表相应的文档数，它们的含义如表 4 所示。

表 4 二值联表

	真正属于此类	真正不属于此类
判断属于此类	a	b
判断不属于此类	c	d

5.4 实验结果

表 5 总结了四种方法在所选数据集上的分类准确率和召回率，从总体上看，本文方法>IG>CHI>MI。由于本文方法使用了优化的 RBF 网络对特征进行优选，使得选择出的特征子集较优秀，所以效果最佳；由于 IG 方法受样本分布影响，在样本分布不均匀的情况下，它的效果就会大大降低，但从整体上看本文所选样本分布相对均匀，只有极个别相差较大，所以总体效果次之；由于 MI 方法仅考虑了特征发生的概率，而 CHI 方法同时考虑了特征存在与不存在时的情况，所以 CHI 方法比 MI 方法效果要好。总的来说，本文所提的方法是有效的，在文本挖掘中有一定的实用价值。

本文首先提出了一个自适应量子粒子群优化算法，用于训练 RBF 网络的基函数中心和宽度，并结合最小二乘法计算网络权值，对 RBF 网络的泛化能力进行改进。然后把该 RBF 神经网络用于特征选择。实验结果表明，采用自适应量子粒子群优化算法获得的 RBF 网络模型不但具有很强的泛化能力，而且具有良好的稳定性，能够选择出较优秀的特征子集。特征选择方法与 3 种经典特征选择方法“信息增益”和“统计量”以及“互信息”相比有较高的准确率和召回率，为后续的知识发现算法减少了时间与空间复杂性，从而使得本文方法在文本分类中有一定的使用价值。

表 5 实验结果

	本文方法		IG	
类别	准确率/%	召回率/%	准确率/%	召回率/%
经济	90.23	91.21	82.52	80.83
体育	88.93	91.47	83.88	82.93
计算机	90.56	90.31	87.64	88.43
政治	87.19	88.41	78.78	84.29
农业	89.90	89.52	83.27	89.67
环境	90.45	89.65	81.67	86.42
艺术	89.37	89.96	80.55	85.81
太空	87.89	91.67	82.46	87.47
历史	90.78	89.39	80.33	87.39
军事	88.69	89.09	75.53	79.73
平均率	89.40	90.07	81.66	85.30
	CHI		MI	
类别	准确率/%	召回率/%	准确率/%	召回率/%
经济	79.31	87.67	75.63	76.99
体育	81.71	85.60	79.54	80.78
计算机	82.41	83.51	80.71	77.91
政治	83.29	78.80	79.99	80.72
农业	79.56	77.23	72.48	79.45
环境	81.93	86.56	76.42	80.13
艺术	82.51	82.78	80.51	81.81
太空	80.84	79.23	78.57	78.47
历史	78.34	80.42	77.45	81.92
军事	60.94	87.67	63.67	74.71
平均	79.08	82.95	76.50	79.29

参考文献

[1] DELGADO M, MARTIN-BAUTISTA M J, SANCHEZ D, et al. Mining text data: special features and patterns [C]// In Proceedings of ESF Exploratory Workshop. London: U. K, Sept, 2002:32-38.

[2] 朱颢东，钟勇.一种新的基于多启发式的特征选择算法[J].计算机应用，2009，29(3)：849-851.

[3] FRIEDMAN N, GEIGER D, GOLDSZMIDT M. Bayesian network classifiers[J]. machineIearning, 1997, 29(2):131-163.

[4] 张海龙，王莲芝.自动文本分类特征选择方法研究[J].计算机工程与设计，2006，27(20)：3838-3841.

[5] 蒋华刚，吴耿锋.基于人工免疫原理的 RBF 网络预测模型[J].计算机工程，2008，34(2)：202-205.

[6] 颜声远，于晓洋，张志俭，等.基于 RBF 网络的显示设计主观评价方法[J].哈尔滨工程大学学报，2007，28(10)：1150-1155.

[7] 臧小刚，宫新保，常成，等.一种基于免疫系统的 RBF 网络在线训练方法[J].电子学报，2008，36(7)：1396-1400.

[8] 刘鑫朝，颜宏文.一种改进的粒子群优化 RBF 网络学习算法[J].计算机技术与发展，2006，16(2)：185-187.

[9] 陈伟，冯斌，孙俊.基于 QPSO 算法的 RBF 神经网络参数优化优化仿真研究[J].计算机应用，2006，26(8)：19-28.

[10] SUN Jun, FENG Bin, XU Wen Bo. Particle swarm pptimization with particles having quantum behavior [A] Proceeding of 2004 Congress on Evolutionary Computation[C]. Piscataway CA: IEEE Press, 2004:325-330.

(收稿日期：2009-04-08)