

结合类内集中度和优化 RBF 神经网络的特征选择

朱 颢 东, 李 红 婵

(郑州轻工业学院 计算机与通信工程学院, 河南 郑州 450002)

摘 要: 特征选择是文本分类的一个核心研究课题. 首先提出了优化的文档频和类内集中度, 紧接着提出了自适应量子粒子群优化算法并用于训练 RBF 网络的基函数中心和宽度, 而且还结合最小二乘法计算网络权值, 对 RBF 神经网络进行了优化, 最后提出了一个综合性特征选择方法. 该综合性方法首先使用类内集中度过滤掉一些词条以降低文本特征空间的稀疏性, 然后再利用优化的 RBF 网络对特征进行优选. 实验结果表明此种特征选择方法有较好的准确率和召回率.

关键词: 特征选择; 文本分类; 量子粒子群优化; RBF 神经网络; 最小二乘法

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2011)02-0145-04

Feature Selection Combined Classificatory Concentration with Improved RBF Neural Network

ZHU Hao-dong, LI Hong-chan

(School of Computer and Communication Engineering,
Zhengzhou University of Light Industry, Zhengzhou 450002, China)

Abstract: Feature selection is the core research topic in text categorization. Firstly, an optimized document frequency and classificatory concentration were presented. And then, an adaptive quantum-behaved particle swarm optimization (AQPSO) algorithm was provided to improve the performance of radial basis function (RBF) neural network. By applying AQPSO algorithm to train the central position and width of the basis function adopted in the RBF neural network, and computing the weights of the network with least-square method (LSM), the RBF neural network was improved. Finally, a combined feature selection method was proposed. The combined feature selection method firstly uses classificatory concentration to filter out some terms to reduce the sparsity of feature spaces, and then employs the improved RBF neural network to select more outstanding feature subset. The experimental results show that the combined method is excellent in accuracy rate and recall rate.

Key words: feature selection; text categorization; QPSO; RBF neural network; least-square method (LSM)

1 引言

文本分类系统通常采用特征集来表示待学习的文档, 这使得特征向量的维数非常大, 有时会达到数十万维. 如此高维的特征对后续的分类过程未必全是重要的、有益的, 而且高维的特征可能会大大增加分类的计算开销, 使整个处理过程的效率非常低下, 而且可能产生与小得多的特征子集相似的分类结

果^[1]. 因此, 必须在保持原文含义的基础上, 找出最能反映文本内容, 又比较简洁的特征向量. 特征选择就是为解决上述问题而产生的一个关键步骤^[2].

常用的文本特征选择方法有信息增益、词频、文档频等^[3-4]. 词频方法选择特征时仅考虑特征在文档集中出现的次数. 由于仅选择出现频繁的词作为特征而忽略了出现频率较低的词. 文档频方法选择特征时仅考虑特征所在的文当数. 其他只考虑特征词

在文档中出现与否,并不考虑特征在文档中出现的次数.这样就导致该方法所选择的特征不具代表性.

通过分析词频方法和文档频方法发现,这两种方法具有互补性.因此,把这两种方法结合起来使用可以获得较好的效果,为此文中提出了一个优化的文档频——最小词频的文档频.

2 类内集中度

定义 1 最小词频的文档频:特征 f_i 的最小词频的文档频是指在类别 c_j 的文本训练集中出现特征 f_i 的次数达到一定数目(用 n 表示)的文档数,记为 $\text{Improved } DF_n(f_i, c_j)$. 其中 n 为特征词 f_i 在类别 c_j 的训练集的文档中至少出现的次数.

如果一个特征对某个类别贡献较大,那么该特征应该集中出现在该类中,而不是分散地出现在各类文档中.为此,又定义了类内集中度,用于表现特征在类别中的集中程度.

定义 2 类内集中度:表示特征 f_i 在类别 c_j 中的集中程度,用 $\text{Class-concentration}(f_i, c_j)$ 表示.由于一个类别的特征词有多个,因此可用以下公式来表示类内集中度:

$$\text{Class-concentration}(f_i, c_j) = \frac{\sum_{k=1 \text{ 且 } k \neq j}^m \left(\frac{\text{Improved } DF_n(f_i, c_j) - \text{Improved } DF_n(f_i, c_k)}{\sum_i \text{Improved } DF_n(f_i, c_j)} \right)^2}{1} \quad (1)$$

其中, m 为类别的个数, $\text{Improved } DF_n(f_i, c_j)$ 定义 1 所定义.经分析可知, $\text{Class-concentration}(f_i, c_j)$ 不但考虑了特征出现的文档数,而且还考虑了特征在文档中出现的次数,把文档频和词频进行了有机的结合. $\text{Class-concentration}(f_i, c_j)$ 越大则表明特征 f_i 在类别 c_j 中的集中程度越大,那么该特征的分类能力也就越强,即该特征也就越重要.

3 RBF 神经网络

RBF 网络是一种 3 层前馈神经网络^[5].它是由输入层、隐含层和输出层组成.网络输出层的输出计算公式为:

$$y_i = \sum_{k=1}^N w_{ik} \varphi_k(x, c_k) = \sum_{k=1}^N w_{ik} \varphi_k(\|x - c_k\|_2) \quad (2)$$

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

$$\varphi(x) = \exp(-x^2/\sigma) \quad (3)$$

$\|\cdot\|_2$ 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; $c_k \in R^{n \times l}$ 表示输入向量的 RBF 网络的中心.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得距离.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].

由式(2)、(3)可见,如果知道了 RBF 网络的隐层节点数 N 、中心 c_k 、宽度 σ , 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 、 σ , RBF 网络模型也建立好了.而对 RBF 网络的隐层节点个数,文中采用了 SOM(self organization mapping) 网络的聚类算法来确定;对中心 c_k 和宽度 σ 则由下面提出的自适应量子粒子群优化算法来确定.

4 自适应量子粒子群优化算法

4.1 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^[6].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^[7]:

$$p = a \times p_b(i) + (1-a) \times g_b \quad (4)$$

$$m_b = \frac{1}{N} \sum_{i=1}^N p_b(i) \quad (5)$$

$$b = 1 - i_t / i_{t_{\max}} \times 0.5 \quad (6)$$

$$p_{\text{pos}} = p \pm b \times |m_b - p_{\text{pos}}| \times \ln \frac{1}{u} \quad (7)$$

其中, p 为 p_b 与 g_b 之间的随机位置, p_b 为粒子的最佳位置, g_b 为粒子群目前的最佳位置; 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 $0 \sim 1$ 之间的随机数,当 $u \geq 0.5$ 时,式(7)取“—”号;当 $0 < u < 0.5$ 时,式(7)取“+”号.

4.2 自适应量子粒子群优化算法(AQPSO)

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.在文献[7]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.从式(6)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,文中做出如下改进:

$$h = f_{\text{fit}2} / f_{\text{fitness}}(i) \quad (8)$$

$$\begin{cases} \text{若 } h < 0.5, b = 2 \times h \\ \text{若 } h \geq 0.5, b = 1 + h \end{cases}$$

其中 $f_{\text{fit}2}$ 为上一代群体获得最佳位置 g_b 的适应度; $f_{\text{fitness}}(i)$ 为第 i 个粒子的当前适应度. h 为两者的比值, h 越小,粒子越远离粒子群的当前最佳位置; h 越大,粒子越靠近粒子群的当前最佳位置.文中以 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; $h \geq 0.5$,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 > 1.8 的概率,使它尽量发散,扩大搜索范围.

5 AQPSO 算法对 RBF 网络的优化

用 AQPSO 算法训练 RBF 神经网络时,首先要向向量形式表示 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 $D = 2m$,每个粒子用一个 $2m$ 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

$c_1, c_2, \dots, c_m$	$\delta_1, \delta_2, \dots, \delta_m$
------------------------	---------------------------------------

表 2 粒子群编码格式

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

计算粒子群体的最佳位置需要比较粒子的适应度,文中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^[8]:

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

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:

Step1: 初始化粒子群体 POP, 每个粒子的最佳位置 $p_b(0) = \emptyset$, 粒子群最佳位置 $g_b = \emptyset$, 粒子的适应度 $f_{\text{fitness}}(0) = 0$, 当前粒子群的最佳适应度 $f_{b \text{ fit}1} = 0$, 上一代粒子群的最佳适应度 $f_{b \text{ fit}2} = 0$ 和预设精度 $\epsilon = 0.09$, 最大迭代次数 $i_{\text{max}} = 200, i = 1$;

Step2: 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 $f_{\text{fitness}}(i)$ 和整个粒子群体的适应度 $f_{b \text{ fit}1}$, 如果 $f_{\text{fitness}}(i) < f_{b \text{ fit}1}$, 则更新粒子 i 最佳位置 $p_b(i)$;

Step3: 判断所有粒子是否完成搜索, 是则转 Step4, 否则返回 Step2;

Step4: 比较当前群体的最佳适应度 $f_{b \text{ fit}1}$ 和上一代群体的最佳适应度 $f_{b \text{ fit}2}$, 若 $f_{b \text{ fit}1} < f_{b \text{ fit}2}$, 则更新粒子群最优位置 g_b 和粒子群的最佳适应度 $f_{b \text{ fit}1}$;

Step5: 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 Step8, 否则转 Step6;

Step6: $i = i + 1$, 如果 $i \geq i_{\text{max}}$, 则转 Step8, 否则返回 Step7;

Step7: 根据式(4)~(9)更新每个粒子的位置, 生成新的粒子群, 返回 Step2;

Step8: 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.

6 特征选择方法

文中特征选择方法结合了两种方法: 类内集中度和基于 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; 然后利用类内集中度过滤掉一些词条(令

式(1)内的最小词频数 $n=3$, 阈值 $m=0.085$, 这时的特征集为初始特征集), 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. 最后用文中优化的 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, 每个初始特征(经过类内集中度筛选后所得的特征)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.

7 实验结果

文中选用的语料库是复旦大学中文文本分类语料库, 其免费下载网址为: <http://www.nlp.org.cn/categories/>. 该语料库包含训练文档集和测试文档集两个部分, 每个部分都包含 20 个类别目录, 每个类别目录存放类别相同的文本, 每个文件都有唯一的编号, 不同文件的编号是不相同的. 实验中只选择前 10 个较大类的部分文档并再从人民网上下载一些相应类别的文档对其进行补充.

文中进行了 10 次实验, 取这 10 次实验结果的平均值作为最终的实验结果, 具体如图 1 和图 2 所示. 图 1 表示四种方法准确率的仿真对比结果. 图 2 表示四种方法召回率的仿真对比结果.

图 1 和图 2 表明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, 文中方法 $> IG > CHI > MI$. 由于文中方法首先用类内集中度过滤掉一些词条以降低文本特征空间的稀疏性, 然后使用优化的 RBF 网络对特征进行优选以选择出较优秀的特征子集, 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, 但从整体上看文中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.

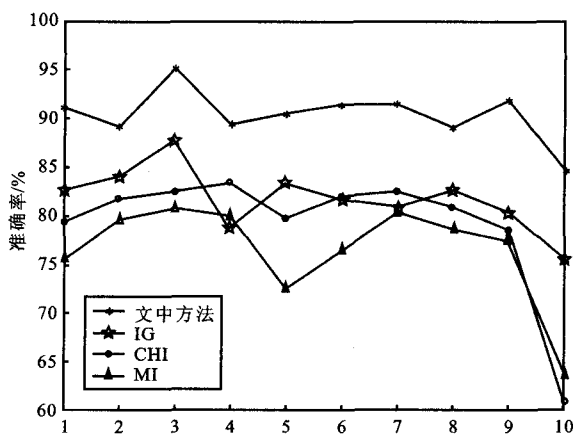


图1 准确率对比结果

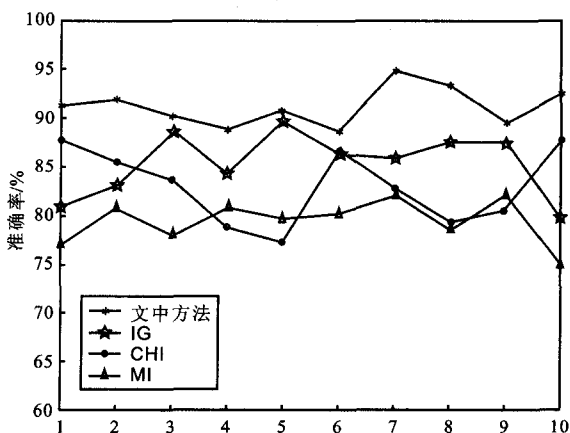


图2 召回率对比结果

8 结束语

文中给出了优化的文档频和类内集中度, 并用类内集中度过滤掉一些词条以降低文本矩阵的稀疏性, 提出了自适应量子粒子群优化算法, 用于训练 RBF 网络^[9]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 神经网络进行优化, 最后用优化的 RBF 网络再一次进行特征优选以选择出较优秀的特征子集. 文中方法不但降低了向量空间的维数, 而且还使选出的特征子集具有较好的代表性. 实验表明, 该综合方法同三种经典特征选择方法“信息增益”和“ χ^2 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.

参考文献:

- [1] Delgado M, Martin Bautista M J, Sanchez D, et al. Mining text data: special features and patterns [C]// Proceedings of ESF Exploratory Workshop. London: U K, 2002: 32-38. (下转第 152 页)

事实上也确实如此. 具体攻击过程中, 每位只需进行一次实验即可猜测出对应位的密钥值.

重复上述实验, 即可得出所有的密钥位.

由实验结果可知, 通过文中所述的 DPA 攻击方法, 可以得出 ECC 所采用的密钥值, 也证明了 Montgomery Ladder 算法不能抗 DPA 攻击.

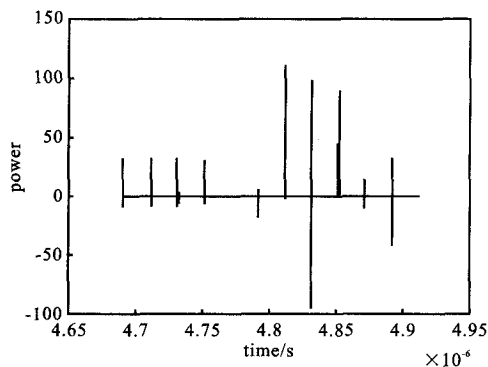


图2 假设 $k_{27}=1$, 密钥位猜测正确, 差分功耗曲线出现尖峰

6 结束语

文中通过对目前常用的 ECC 点乘算法进行分析, 首先用 Verilog HDL 描述出其 RTL 级代码, 并用 Chartered CMOS 工艺综合得到其电路网表, 以便更精确采集到点乘运算所产生的功耗曲线. 接下来用 ZEMD 功耗分析攻击方法, 成功的获得 ECC 加密运算所用到的密钥值. 实验结果显示, Montgomery Ladder 算法无法抗 DPA 攻击, 因此在实际应用中还要采取一些措施以增强其安全性. 抗 DPA 攻击的对策研究成为下一步工作的主要内容.

参考文献:

- [1] Kocher, Jaffe P J, Jun B. Differential power analysis [J]. CRYPTO'99, 1999(LNCS 1666): 388-397.
- [2] 赵彦光, 白国强, 陈弘毅. 一种针对特征 2 域椭圆曲线密码芯片的差分功耗分析[J]. 微电子学与计算机, 2005, 22(12): 78-81.
- [3] 刘鸣, 陈弘毅, 白国强. 功耗分析研究平台及其应用[J]. 微电子学与计算机, 2005, 22(7): 134-138.
- [4] 孙红霞, 谭征. ECC 密码技术及其优越性[J]. 信息技术与信息化, 2005(4): 87-88.
- [5] Darrel Hankerson, Alfred Menezes, Scott Vanstone. 椭圆曲线密码学导论[M]. 张焕国, 译. 北京: 电子工业出版社, 2005.
- [6] Lopez J, Dahab R. Fast multiplication on elliptic curves over $GF(2^m)$ without precomputation [J]. CHES 1999, 1999(LNCS 1717): 316-327.
- [7] Itoh K, Izu T, Takenaka M. Address-bit differential power analysis of cryptographic schemes OK-ECDH and OK-ECDHSA[J]. CHES 2000, 2000(LNCS 2523): 129-143.
- [8] Thomas S Messerges, Ezzy A Dabbish, Robert H Sloan. Power analysis attacks of modular exponentiation in smart-cards[J]. CHES 1999, 1999(LNCS1717): 144-157.

作者简介:

邓秋成 男, (1986-), 硕士研究生. 研究方向为集成电路设计、信息安全.

白雪飞 男, (1977-), 讲师. 研究方向为集成电路设计、信息安全.

郭立 男, (1964-), 教授, 博士生导师. 研究方向为集成电路设计、数字信号处理、信息安全.

王尧 男, (1963-), 硕士研究生. 研究方向为集成电路设计、信息安全.

(上接第 148 页)

- [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]. Machine Learning, 1997, 29(2): 131-163.
- [4] 陈平, 刘晓霞, 李亚军. 文本分类中改进型互信息特征选择的研究[J]. 微电子学与计算机, 2008, 25(6): 194-197.
- [5] 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.
- [6] 刘成, 肖扬. 粒子群优化算法在多用户检测中的应用[J]. 中国铁道科学, 2006, 27(4): 129-132.
- [7] Sun Jun, Feng Bin, Xu Wenbo. Particle swarm optimization with particles having quantum behavior[C]//Proceeding of 2004 Congress on Evolutionary Computation. Piscataway CA: IEEE Press, 2004: 325-3.

tion with particles having quantum behavior[C]//Proceeding of 2004 Congress on Evolutionary Computation. Piscataway CA: IEEE Press, 2004: 325-3.

- [8] 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.

- [9] 邢桂华, 朱庆保. 基于模拟退火遗传算法的 RBF 网络的优化[J]. 微电子学与计算机, 2005, 22(7): 174-177.

作者简介:

朱颢东 男, (1980-), 博士. 研究方向为文本挖掘、智能信息处理、计算智能.

李红婵 女, (1983-), 硕士. 研究方向为智能信息处理.