基于改进 QPSO 和 RBF 神经网络的文本分类方法①

李滨旭[1], 姚姜虹[2]

[1](东北石油大学　计算机与信息技术学院, 大庆 163318)
[2](大庆市油田信息技术公司物联网分公司, 大庆 163318)

摘　要: 为提高文本分类的准确性, 本文提出了一种基于量子 PSO 和 RBF 神经网络的新的文本分类方法. 首先建立描述样本类别的关键词集合, 并采用模糊向量空间模型建立每类样本的特征向量, 然后采用 RBF 神经网络实施文本自动分类, 采用改进的量子PSO优化RBF神经网络的参数, 以提高其逼近能力. 选取中国期刊网的部分文献作为实验数据, 实验结果说明本文所提出方法的分类精准度与其他同类方法相比有明显的提高.

关键词: 文本分类; 量子 PSO; RBF 神经网络; 算法设计

Document Classification Based on Improved QPSO and RBF Neural Networks

LI Bin-Xu[1], YAO Jiang-Hong[2]

[1](School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318,China)
[2](Daqing Petroleum Information Technology IoT Branch Company, Daqing 163318, China)

Abstract: To enhance the accuracy of the text classification, a new method based on quantum PSO and RBF neural network is proposed. Firstly, it establishes the key words set to describe the classification of the samples, and uses fuzzy vector space model to build the feature vectors of every kind of sample, then automatically classifies the texts by RBF neural network, optimizes the parameters of RBF neural network by improved quantum PSO to enhance its approximation capability. The new method is proved by the classification of some documents in China periodical document database. The experiment shows that this method makes significant improvements in classification accuracy compared to other methods.

Key words: text classification; quantum PSO; RBF neural network; algorithm design

1　引言

文本分类自提出以来得到了广泛的关注, 经过了不断的发展, 已成为管理文本数据的一种核心方式[1]. 作为数据挖掘的一个重要分支, 其分类过程可描述为一种学习过程. 首先根据已有的训练文本集找到文本特征与文本分类之间的关系, 形成分类器, 再利用分类器对新的文本进行分类. 目前, 学者们对于文本分类的研究已经取得许多进展, 主要方法有 K 近邻分类法(KNN)、普通贝叶斯算法(Bayes)、神经网络等[2]. K 近邻分类方法可有效避免样本不平衡的问题, 但是在样本维数较高时, 其计算复杂度较高[3]; 贝叶斯算法较简单, 易实现, 但受特征向量的影响其分类精度不够高; RBF 神经网络是一种局部逼近的神经网络, 其优化过程可看作在高维空间中的曲面拟合问题, 被广泛应用于函数曲线逼近和模式分类问题[4]. 粒子群优化(Particle Swarm Optimization, PSO)算法是一种全局优化算法, 该算法具有容易实现、收敛速度快等优点, 目前已被广泛应用到神经网络训练中[5], 粒子群算法用于神经网络优化主要包括三个方面, 一是用于网络学习(也称网络训练), 即优化网络全连接结构下的各层连接权值, 二是优化网络的拓扑结构, 三是在优化网络拓扑结构的同时进行网络学习. QPSO 算法是由 PSO 算法发展而来的, 在收敛速度上有较大的提高[6,7]. 本文提出了一种基于量子势阱中心进行改进的量子粒子

①基金项目:东北石油大学研究生创新科研项目(YJSCX2016-030NEPU)
收稿时间:2016-01-07;收到修改稿时间:2016-02-26　[doi: 10.15888/j.cnki.csa.005340]

证明了方法的可行性. 关键词库的结合大大提高了信息抽取算法的准确性和通用性, 基于 Web 信息抽取的群优化算法(IQPSO), 利用其优化 RBF 网络中的参数, 建立IQPSO-RBF网络模型对文本进行分类, 实验部分混合交通出行方案生成与表示系统的成功实验也证明了本文提出的 Web 信息抽取算法的实用性.

2 改进的量子行为粒子群算法

PSO 是由 Eberhart 和 Kennedy 等人于 1995 年提出的[8], 它是一种基于种群搜索的自适应进化计算技术. 自提出以来, 众学者们对其进行了深入的研究, 主要从参数、进化方程以及与其他算法的融合几个方面入手. 现有的 QPSO 的迭代方程为

$$x_{k+1} = P \pm \alpha | x_k - P | \ln(1/u) \tag{1}$$

QPSO 迭代式为[9]

$$x_{id}(t+1) = P_{id}(t) \pm \alpha \ln(1/u) | x_{id}(t) - C_d(t) | \tag{2}$$

式中, $\mathbf{P}_i$ 为第 i 个粒子的自身最优位置, $\boldsymbol{C}(t)$ 为所有 M 个粒子自身最优位置的算术平均值, 即

$$\boldsymbol{C}(t) = \frac{1}{M}\left(\sum_{i=1}^{M} p_{i1}(t), \sum_{i=1}^{M} p_{i2}^{L}(t), \cdots, \sum_{i=1}^{M} p_{in}^{L}(t)\right) \tag{3}$$

本迭代在后期阶段中, 作为优化路标的势阱中心不具有较强的引领作用. 由此, 本文拟从当代构造的最优解候选集中, 采用轮盘赌注的方法随机选取每步迭代的势阱中心粒子.

最优解候选集由根据当代种群适应度择优选取前 K_{best} 个粒子组成. 众所周知, 在算法初期, 重点在于全局探索, 而在算法后期, 重点在于局部开发. 因此, 为平衡算法的探索和开发能力, 我们使候选集中粒子的数目 K_{best} 随迭代步数单调减小. 具体如下式所示.

$$K_{best} = \left\lfloor popsize \times 0.1^{t/Max\_N} \right\rfloor \tag{4}$$

式中 $popsize$ 为种群规模, Max_N 为限定步数, t 为当前步数.

另外, 为增强种群多样性, 使算法避免早熟收敛, 本文基于当代种群所有粒子的平均值(中心粒子)设计了一个新的位置更新式. 该式的作用为, 使粒子等概率地向着中心粒子或偏离中心粒子移动.

记第 t 代用轮盘赌选择的粒子为 $\mathbf{X}_{best}(t)$, 种群平均粒子为 $\mathbf{X}_{mean}(t) = \frac{1}{popsize}\sum_{i=1}^{popsize} \mathbf{X}_i(t)$, 本文采用随机选择如下两式之一更新粒子位置.

$$X_i^d(t+1) = X_{best}^d(t) \pm \alpha | X_i^d(t) - X_{best}^d(t) | \ln(\frac{1}{u}) \tag{5}$$

$$X_i^d(t+1) = X_i^d(t) \pm \alpha | X_i^d(t) - X_{mean}^d(t) | \ln(\frac{1}{u}) \tag{6}$$

由式(4)-(6)可知, 在本文提出的 IQPSO 中,除了种群规模、变量维数、迭代步数等所有智能优化算法都需要事先设定的共性参数外,体现算法自身特性的控制参数, 并没有增加, 仍然只有一个 α.

值得指出 IQPSO 与现有 QPSO 有两点不同. 第一, 在 IQPSO 中, 每步迭代用作势阱中心的粒子来自于由前 K_{best} 个适应度最大粒子组成的候选集, 具体采用哪个粒子由轮盘赌策略决定; 而 QPSO 自始至终分别采用自身最优粒子和自身最优粒子的算术平均作为势阱中心. 第二, 关于粒子位置的更新, IQPSO 除采用由量子势阱建模得到的更新式外, 还采用了另一个类似的更新式, 每步迭代, 两个公式等概率随机选择使用.

3 基于IQPSO的RBF神经网络分类模型

3.1 RBF 神经网络模型

径向基神经网络模型由输入层、隐层和输出层组成, 网络结构如图 1 所示[10]. 其中, 输入层 n 个节点; 隐层 h 个节点; 输出层 m 个节点. $(x_1, x_2, \cdots, x_n)$ 为输入模式特征向量; $(y_1, y_2, \cdots, y_m)$ 为模式输出向量; $W_k = (w_{1k}, w_{2k}, \cdots, w_{hk})^T, (k = 1,2,\cdots,m)$ 为输出节点 k 的权值向量. RBF 采用高斯函数, 隐层第 i 节点的输出为

$$q_i = \exp(-(\frac{X - c_i}{\sigma_i})^2) \tag{7}$$

式中, X : n 维输入向量; c_i :第 i 个隐节点的中心; σ_i :第 i 个隐节点的宽度. $i = 1,2,\cdots,h$. 输出层第 k 个节点的输出为隐节点输出的线性组合:

$$y_k = \sum_{i=1}^{h} w_{ik} q_i - \theta_k \tag{8}$$

式中, $w_{ik} : q_i \to y_k$ 的连接权; θ_k :第 k 个输出节点的阈值.

关于 RBF 结构参数的确定方法:输入节点个数为样本维数, 需要根据实际问题确定, 隐层节点个数为样本类数, 也需要根据实际问题确定, 对于分类问题, 输出节点取一个即可. 关于RBF网络参数的确定方法: 输入层权值恒为 1, 隐层中心初值取各类样本均值, 方

差初值取 $\frac{d_{\max}}{\sqrt{M}}$，其中 M 为样本类数，$d_{\max}$ 为样本两两之间的最大距离，输出层权值取(-1,1)之间的随机数.

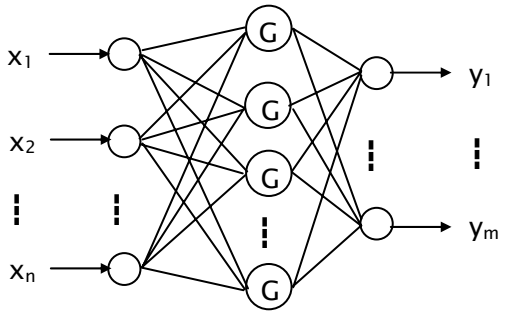

图 1　径向基神经网络模型

3.2 Web 信息抽取

本模块算法流程如图 2 所示.

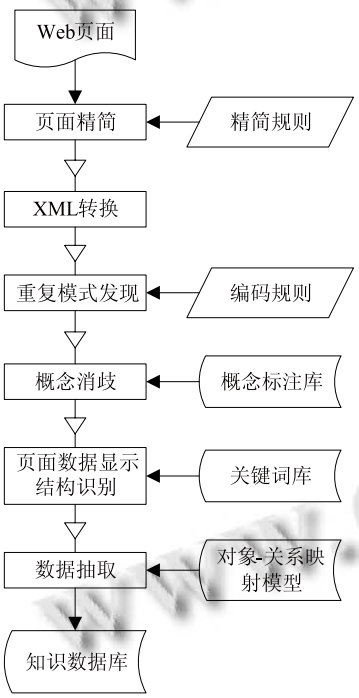

图 2　Web 信息抽取流程

3.2.1 页面精简

普通网页常常包含很多 Header 部页面属性信息、脚本、样式、注释、图片、隐含数据、空格、标签属性设置及一些无用标签等，这些信息中不含有集中式数据，对造成后续处理速度缓慢，甚至使后续处理无法进行，需要首先进行页面精简，去掉这些冗余信息. 本系统采取采用正则表达式技术进行如下页面精简操作:

① 清除 body 以外的部分;

② 清除文档中的脚本(<script 脚本内容</script>)、样式(<style 样式内容</style>)、注释(<!--注释内容-->)、隐含内容(<input type=”hidden” 隐含内容>)、图片内容(<img 图片内容>);

③ 清除文档中没有实际内容的标签对(只含空格、换行符等)(递归清除);

④ 将连续多个“ ”和“ ”替换成一个空格“ ”;

⑤ 清除标签的属性信息.

3.2.2 XML 转换

由于 HTML 语法的随意性，即使经过页面精简，仍无法保证 HTML 文档的结构特性. 而 XML 是一种结构化的自解释语言，更方便于进行重复模式发现，且在数据抽取过程中采用了 XML 的对象-关系映射技术，需要将 HTML 文档转换成 XML 文档.

本系统采用开源的 Jtidy 工具，实现 HTML 文档到 XML 文档的转换[11].

3.2.3 重复模式发现

数据密集型 Web 页面的一个显著特点是数据显示区域(绝大部分情况是列表或表格形式)具有很强的重复模式，针对这一特点，可以通过重复模式的发现，很方便的确定页面数据显示区域的结构.

本系统采用基于 PAT-array 的算法实现快速的文档内重复模式的发现. 具体步骤如下:

① 令牌翻译: 对 HTML 中与数据显示相关的标签进行编码，将转换得到的 XML 文档翻译成二进制字符串;

② PAT 数组构造: 罗列二进制字符串的所有半串(从每个编码到结束位置构成一个半串)，按序排列后得到每个半串起始位置序号构成 PAT 数组;

③ 候选重复模式发现: 使用栈操作，搜索得到所有半串的共同前缀即为候选重复模式;

④ 最佳重复模式确定: 根据最优化标准从候选重复模式中确定出最佳重复模式.

3.2.4 概念消歧

单纯的重复模式发现算法只能得到笼统的数据显示结构，无法区分真正的数据及其语义(标题). 本系统采用基于本体的关键词库从重复模式中区分出标题项和数据项，最终确定准确的数据显示结构.

对于自然语言表示的 Web 文档, 其中存在大量同义的词汇, 在进行标题识别前需要进行概念消歧处理, 利用概念标注库, 将特定领域的同义词汇转换为关键词库中的本体词.

3.2.5 页面数据显示结构识别

本系统采用 XML 的对象-关系映射技术实现数据抽取, 页面数据显示结构的识别即为 XML 文档对象模型(DOM)的确定. 步骤如下:

① 标题定位: 使用关键词库中特定领域的本体词集合, 对页面中符合重复模式的数据进行搜索和定位, 确定出其中的标题项;

② 标题-数据映射关系识别: 根据确定出来的标题项集合的相对关系及与重复模式中其他数据项的相对关系, 确定出各个标题项与数据项的映射关系;

③ DOM 树生成: 根据重复模式及确定出的各个标题项与数据项的映射关系, 生成对应的 DOM 树.

对于如下的 xml 文档:

```
<?xml version="1.0" encoding=   "GB2312"?>
<table>
  <tr>
    <td>车次</td>
    <td>1019</td>
    </tr>
  <tr>
    <td>始发站</td>
    <td>合肥</td>
    <td>终点站</td>
    <td>东莞东</td>
  </tr>
  ……
</table>
```

4 结语

数据密集型页面往往由 Web 站点根据用户的查询请求动态生成, 从同一站点能得到大量同类型的动态页面. 据此, 系统以知识数据库为基础, 采用 Web 站点配置方式, 根据 Web 站点响应查询请求方式, 人工配置含特定知识的 Web 站点信息及其动态页面 URL 生成规则. 用知识数据库中现有知识作为查询参数, 生成相关 Web 站点的动态 URL, 通过 HTTP 协议自动获取相关 Web 页面.

数据密集型页面往往由 Web 站点根据用户的查询请求动态生成, 从同一站点能得到大量同类型的动态页面. 据此, 系统以知识数据库为基础, 采用 Web 站点配置方式, 根据 Web 站点响应查询请求方式, 人工配置含特定知识的 Web 站点信息及其动态页面 URL 生成规则. 用知识数据库中现有知识作为查询参数, 生成相关 Web 站点的动态 URL, 通过 HTTP 协议自动获取相关 Web 页面.

参考文献

1 石东源,卢炎生,王星华,段献忠.SVG 及其在电力系统软件图形化中的应用初探.继电器, 2004,32(16):37−40.

2 刘崇茹,孙宏斌,张伯明,董越,辛耀中.基于 CIM XML 电网模型的互操作研究.电力系统自动化,2003,27(14):45−48.

3 朱丽娟,王康元,张洁.基于 SVG 和 Java 的电力系统节点电压可视化.继电器, 2006,34(5):60−61.

4 王志春,杨军,胡桂杰.基于 Surfer Automation 接口的气象等值线图的绘制.内蒙古气象,2006:31−33.

5 陈志波,陆雍森.Surfer 在环境评价和规划中的应用.同济大学学报(自然科学版),2005,33(2):191－195.

6 章坚民,徐爱春,李海翔等.基于 SVG/XML/CIM 的变电站自动化工程配置系统.电力系统自动化,2004,28(14):53−56.

7 Zhang JM, Xu AC, Li HX et al. An automatic engineering configuration system for substation automation based on SVG/XML/CIM. Automation of Electric Power Systems, 2004, 28(14): 53−56.

8 章坚民,楼坚.基于 CIM/SVG 和面向对象的配电单线图自动生成.电力系统自动化,2008,32(22):61−65.

9 吴延霞.Surfer 软件嵌入 VB 编程在激光平地系统中的应用.德州学院学报,2006,12(6):101−102.

10 尼建军,张学宏.Surfer 7.0 嵌入 VB 6.0 编程实现水文数据快速可视化.海洋测绘,2005,25(1):64−65.

11 韩丽娜,石昊苏.利用 Surfer 8.0 绘制地质等值线图.计算机与现代化,2008,11:85−88.

12 张二勇,李云峰,王玮.Surfer 软件绘图接口的开发及应用.地下水,2005,27(3):212−214.