基于多种群协同优化的文本分类规则抽取方法

刘 赫 [1,2] 刘大有 [1,2] 裴志利 [3] 高 滢 [1,2]

摘 要 针对文本分类中的规则抽取问题, 提出一种基于多种群协同优化的文本分类规则抽取方法. 该方法利用信息熵生成初始种群, 采用多种群协同优化方法演化当前种群. 多种群协同优化方法通过种群之间的相互竞争和良种共享机制提高优化方法的效率. 实验结果表明, 本文提出的文本分类规则抽取方法所抽取规则的数量少, 准确率高, 平均长度短; 同时, 本文方法所用的计算时间少, 抽取分类规则的速度快, 适用于大规模数据集.

关键词 规则抽取, 文本分类, 多种群协同优化, 遗传算法, 蚁群算法
中图分类号 TP391

Rule Extraction Approach to Text Categorization Based on Multi-population Collaborative Optimization

LIU He[1,2] LIU Da-You[1,2] PEI Zhi-Li[3] GAO Ying[1,2]

Abstract For the problem of rule extraction in text categorization, a novel rule extraction approach to text categorization based on multi-population collaborative optimization was proposed. Information entropy was applied to generation of initial populations and the multi-population collaborative optimization method was employed to evolve the current population in this proposed approach. The optimization efficiency of this approach was improved by the mutual competition and excellent individuals sharing mechanisms among populations. Experimental results have shown that the number of the rules extracted by this approach is small, and that the accuracy of these rules is high and the average length of them is short. Furthermore, the time of this approach is short and the speed of rule extraction through this approach is high. Therefore, this approach is suitable for large-scale data sets.

Key words Rule extraction, text categorization, multi-population collaborative optimization, genetic algorithm, ant colony algorithm

文本分类 (Text categorization, TC) 作为处理和组织大量文本数据的关键技术, 可在较大程度上解决信息杂乱问题, 方便人们准确定位和分类所需要的信息. 目前, 文本分类作为数据挖掘领域的一个重要分支, 受到了人们的广泛关注, 并且在搜索引擎、信息过滤、信息检索、信息分发、文本数据库和数字化图书馆等领域都有广泛的应用.

文本分类出现于 20 世纪 60 年代初期, 直到 20 世纪 80 年代末, 在文本分类中占主导地位的一直是基于知识工程的分类方法. 知识工程方法一般采用析取范式为每个类别定义逻辑规则, 可以理解为一种简单的自然语言处理方法. 知识工程方法需要手工编写规则或应用其他复杂的自然语言处理技术, 难度非常大, 也非常耗时, 在很多场合显得过于低效而不实用. 20 世纪 90 年代以来, 知识工程方法逐渐被基于机器学习的方法所取代. 机器学习就是用计算机辅助人来学习关于认识世界和改造世界的知识. 机器学习方法能够获得与人类专家相媲美的分类效果, 并且节省人力, 因此比知识工程方法更有吸引力. 目前, 常用的基于机器学习的文本分类方法主要有: k 近邻方法[1], 朴素贝叶斯方法[2], 支持向量机方法[3], Rocchio 方法[4], Boosting 方法[5], 决策树方法[6], 关联规则方法[7], 神经网络方法[8], 最大熵方法[9] 和粗糙集方法[10] 等.

现有的基于机器学习的文本分类方法大部分都基于向量空间模型. 向量空间模型是一种简单且有

收稿日期 2008-05-27 收修改稿日期 2009-04-15
Received May 27, 2008; in revised form April 15, 2009
国家自然科学基金重大项目 (60496321), 国家高技术研究发展计划 (863 计划) (2006AA10Z245, 2006AA10A309), 国家自然科学基金 (60773099, 60573073), 吉林省科技发展计划重大项目 (20020303), 吉林省科技发展计划项目 (20030523), 欧盟项目 TH/Asia Link/010 (111084) 资助
Supported by Key Program National Natural Science Foundation of China (60496321), National High Technology Research and Development Program of China (863 Program) (2006AA10Z245, 2006AA10A309), National Natural Science Foundation of China (60773099, 60573073), Key Program Science and Technology Development Plan of Jilin Province (20020303), Science and Technology Development Plan of Jilin Province (20030523), and European Commission TH/Asia Link/010 (111084)
1. 吉林大学计算机科学与技术学院 长春 130012 2. 吉林大学符号计算与知识工程教育部重点实验室 长春 130012 3. 内蒙古民族大学计算机科学与技术学院 通辽 028043
1. College of Computer Science and Technology, Jilin University, Changchun 130012 2. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012 3. College of Computer Science and Technology, Inner Mongolia University for Nationalities, Tongliao 028043
DOI: 10.3724/SP.J.1004.2009.01334

效的模型, 基于该模型的方法具有很好的分类效果. 然而, 当文本集合新增文本时, 这类方法需要重新生成分类器, 可扩展性差[11]. 而且, 这类方法很难抽取知识, 即易于理解的分类规则[12]. 虽然有一些基于规则的文本分类方法, 比如: 传统的 CN2 方法[13], 基于蚁群优化的 Ant-Miner 方法[14], 基于模糊决策树的 FDT 方法[12], 基于线性支持向量机的 LSVM 方法[15] 和基于递归神经网络的 RNN 方法[16] 等, 但是有些方法抽取易于理解的分类规则仍比较困难. 本文提出了一种基于多种群协同优化 (Multi-population collaborative optimization, MPCO) 的文本分类规则抽取方法, 该方法利用信息熵生成初始种群, 采用多种群协同优化方法演化当前种群. 实验表明, 运用本文方法所抽取的规则数量少, 准确率高, 平均长度短; 同时, 本文方法所用的计算时间少, 抽取分类规则的速度快, 适用于大规模数据集.

1　基于规则抽取的 Web 文本分类

文本挖掘是数据挖掘领域的一个重要分支. 它首先采用文本切分技术对文本的特征进行抽取, 将文本数据转化为能够描述文本内容的结构化数据; 然后利用分类、聚类和关联分析等数据挖掘技术形成结构化文本树, 并在所形成的结构中发现新的概念和获取相应的关系. 以 Web 文本为对象的文本挖掘被称为 Web 文本挖掘. Web 文本挖掘中一个关键分支是 Web 文本分类. 基于规则抽取的 Web 文本分类的基本流程如图 1 所示.

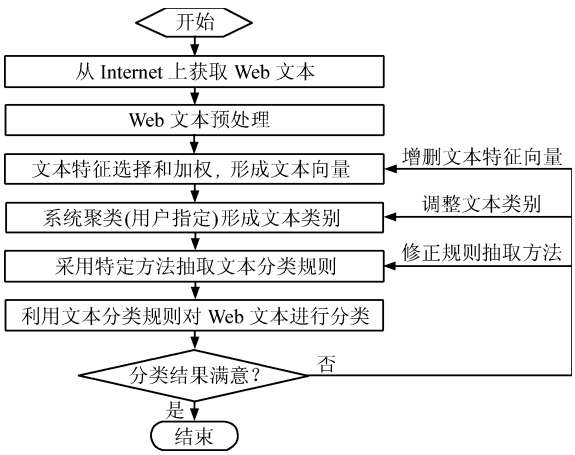


图 1　基于规则抽取的 Web 文本分类的流程

Fig. 1　The flow of web text categorization based on rule extraction

如图 1 所示, 在基于规则抽取的 Web 文本分类的流程中, 最核心的步骤就是采用特定方法抽取文本分类规则, 这也是本文研究的重点. 本文提出了一种基于多种群协同优化的文本分类规则抽取方法. 该方法利用信息熵生成初始种群, 采用多种群协同优化方法演化当前种群, 通过种群之间的相互竞争和良种共享机制提高优化方法的效率. 本文方法试图挖掘出一个能覆盖大多数甚至全部训练样本的分类规则列表. 本文方法的优化框架如图 2 所示.

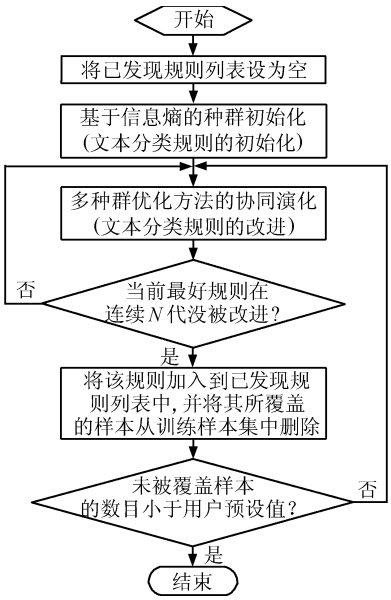


图 2　本文方法的优化框架

Fig. 2　The optimization frame of the approach proposed by this paper

2　基于信息熵的种群初始化

文本分类中的规则抽取就是挖掘蕴含在训练样本集中的分类规则. 本文采用基于信息熵的方法[17] 来生成初始种群, 该方法采用一个数组来表示一条分类规则, 每条分类规则的结构如图 3 所示.

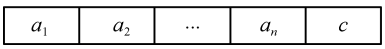


图 3　分类规则的结构

Fig. 3　The structure of categorization rule

在图 3 中, $a_i\,(i = 1, 2, \cdots, n)$ 表示分类规则中的一个属性, c 表示分类规则所对应的类别.

在给出基于信息熵的种群初始化方法之前, 先介绍该方法涉及到的几个概念.

1) 特征. 特征 t_{ij} 表示条件 "$a_i = v_{ij}$", 其中, a_i 代表一条规则的第 i 个属性, v_{ij} 表示属性 a_i 的第 j 个取值. 例如, 假设 "年龄" 为一条规则的第 2 个属性, 有 "少年"、"青年"、"中年" 和 "老年"4 个取值, 则特征 t_{21} 代表条件 "年龄 = 少年", 特征 t_{22} 代表条件 "年龄 = 青年", 特征 t_{23} 代表条件 "年龄 = 中年", 特征 t_{24} 代表条件 "年龄 = 老年".

2) 特征的启发函数. 基于信息熵的种群初始化方法通过一个基于信息熵理论构建而成的启发函数来估算每个特征提高当前规则预测准确率的能力.

定义 1. 假设集合 D 是训练样本集, 其中样本的数目为 l, $d_k \in D\,(k = 1, 2, \cdots, l)$ 是训练集 D 中的一个样本, 那么, 特征 t_{ij} 的信息熵 $H(t_{ij})$ 的计算公式如下:

$$H(t_{ij}) = -\sum_{k=1}^{l}(P(t_{ij}|d_k)\log_2 P(t_{ij}|d_k)) \qquad (1)$$

其中, $P(t_{ij}|d_k)$ 表示特征 t_{ij} 出现在训练样本 d_k 中的概率.

特征的信息熵表示该特征在训练样本集中的分布情况. $H(t_{ij})$ 的值越大, 表示特征 t_{ij} 在训练样本集中分布越均匀, 此时, 特征 t_{ij} 将以较小的概率加入到当前规则中, 反之亦然.

定义 2. 假设 n 为属性的数目, $m_i (i = 1, 2, \cdots, n)$ 为属性 a_i 取值的数目, x_i 表示一个布尔变量, 如果属性 a_i 未被选到当前规则中, x_i 取值为 1, 反之, 取值为 0. 那么, 特征 $t_{ij}\,(j = 1, 2, \cdots, m_i)$ 的启发函数 $E(t_{ij})$ 的计算公式如下:

$$E(t_{ij}) = \frac{\log_2 l - H(t_{ij})}{\sum\limits_{p=1}^{n}\left[x_p \sum\limits_{q=1}^{m_p}(\log_2 l - H(t_{pq}))\right]} \qquad (2)$$

特征的启发函数反映了该特征的平均信息量. $E(t_{ij})$ 的值越大, 表示特征 t_{ij} 与当前分类的相关性越大, 此时, 该特征将以较大的概率加入到当前规则中, 反之亦然.

3) 特征的插入概率. 每个特征按照一定的概率被插入到当前规则中.

定义 3. 假设 μ_{ij} 表示特征 t_{ij} 被选到分类规则中的次数, 那么, 特征 t_{ij} 的插入概率 $I(t_{ij})$ 的计算公式如下:

$$I(t_{ij}) = \frac{x_i \mu_{ij} E(t_{ij})}{\sum\limits_{p=1}^{n}\left[x_p \sum\limits_{q=1}^{m_p}(\mu_{pq} E(t_{pq}))\right]} \qquad (3)$$

特征的插入概率反映了该特征对分类的重要程度. $I(t_{ij})$ 的值越大, 表示特征 t_{ij} 对分类越重要, 此时, 特征 t_{ij} 将以较大的概率被插入到当前规则中, 反之亦然.

4) 特征插入的条件. 如果特征 t_{ij} 能同时满足以下两个条件, 那么该特征将以概率 $I(t_{ij})$ 被插入到当前规则中.

a) 特征 t_{ij} 中的属性 a_i 不在当前规则中;

b) 特征 t_{ij} 加入到当前规则后, 该规则在当前状态所覆盖的训练样本的数目大于或者等于预先设定的分类规则所覆盖的最小样本数目.

5) 特征插入的终止条件. 如果当前状态满足以下条件之一, 那么停止向当前规则中插入特征.

a) 当前规则所包含的特征数目大于或者等于属性的数目;

b) 没有特征满足特征插入条件.

根据上面介绍的概念, 基于信息熵的种群初始化方法可以描述如下:

步骤 1. 生成空规则库;

步骤 2. 生成一条空规则;

步骤 3. 从特征集合中随机选择一个特征 t_{ij};

步骤 4. 如果特征 t_{ij} 满足特征插入的条件, 则将该特征以概率 $I(t_{ij})$ 插入到当前规则中, 否则, 转到步骤 3;

步骤 5. 如果当前状态不满足特征插入的终止条件, 则转到步骤 3;

步骤 6. 为当前规则选择一个类别, 要求该类别能使当前规则最大程度地覆盖训练样本集中的样本;

步骤 7. 将当前规则加入到规则库中, 并将其所覆盖的样本从训练样本集中删除;

步骤 8. 重复步骤 2 ~ 7, 直到训练样本集为空或者规则库中的规则数目大于或者等于预先设定的值.

3 多种群协同优化方法

3.1 多种群协同优化的框架

在自然界中, 不同地域的生物有不同的特点和进化程度, 它们从大自然中争夺资源为己所用. 同时, 这些生物之间又通过信息交换来取长补短并共同进化. 本文提出的多种群协同优化方法就是借鉴自然界的这一现象设计出来的. 多种群协同优化系统的结构如图 4 所示.

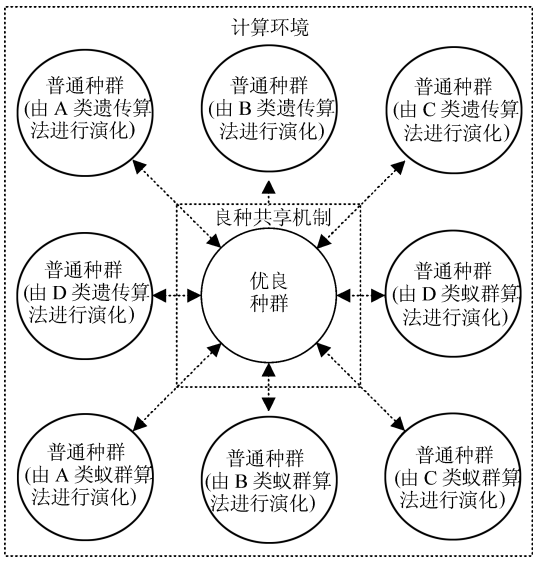


图 4 多种群协同优化系统的结构

Fig. 4 The structure of multi-population collaborative optimization system

从图 4 可以看出, 多种群协同优化系统由计算环境中的多个普通种群和一个优良种群构成. 各个普通种群通过相互竞争机制在计算环境中竞争计算资源. 一旦某个种群获得计算资源, 便进行一次自己的进化过程. 各个普通种群将进化得到的优良个体贡献出来, 组成优良种群. 普通种群可以从优良种群中获取优良个体, 以改善本种群的品质. 各个普通种群的演化通过各种不同的遗传算法和蚁群算法来完成. 各种遗传算法和蚁群算法之间的不同体现在演化机制、优化算子和控制参数等方面的差异. 限于篇幅, 本文不再赘述各种遗传算法和蚁群算法的细节.

多种群协同优化方法的流程如图 5 所示.

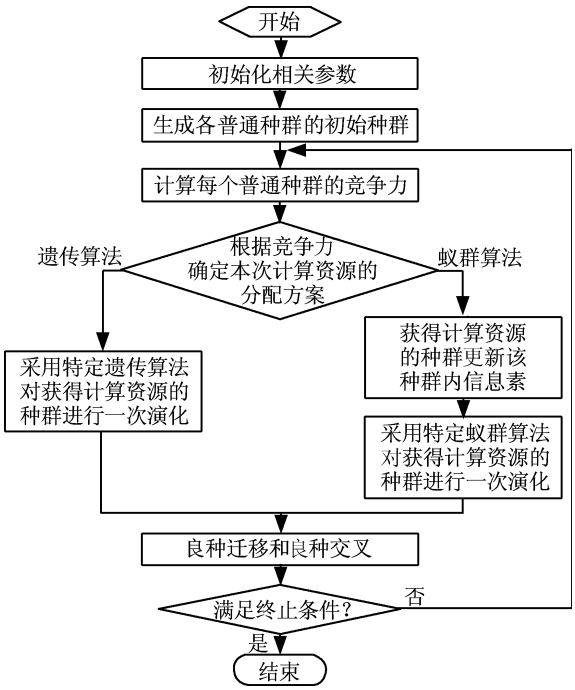


图 5　多种群协同优化方法的流程

Fig. 5　The flow of multi-population collaborative optimization approach

3.2　适应度评价函数

本文从如下 3 个方面设计适应度评价函数.

1) 支持度 (Support). 规则的支持度是指成功匹配规则的样本在整个样本集中所占的比例. 支持度反映了规则在整个样本集中所具有的普遍性.

2) 置信度 (Confidence). 规则的置信度是指规则中由特征属性导致类别属性的可信程度. 置信度反映了不完全知识条件下推理的不确定性. 置信度越大, 说明条件为真时越容易产生相应的结论.

3) 覆盖度 (Coverage). 规则的覆盖度是指同时与规则的特征属性部分和类别属性部分成功匹配的样本在只与规则的类别属性部分成功匹配的样本中所占的比例. 覆盖度反映了结论包含于条件的准确程度. 覆盖度越大, 说明规则越完备.

适应度函数的设计应综合考虑规则的特征属性部分和类别属性部分以及它们与样本的匹配情况. 本文把支持度、置信度和覆盖度综合起来设计适应度评价函数. 在进化过程中, 每条规则依靠自身的适应度函数值与其他规则进行竞争, 只有支持度、置信度和覆盖度都高的规则, 才能在竞争中被保留下来.

定义 4.　假设 N 表示待分类样本的数目, $TN(r)$ 表示成功匹配规则 r 特征属性部分的待分类样本的数目, $CN(r)$ 表示成功匹配规则 r 类别属性部分的待分类样本的数目, $RN(r)$ 表示成功匹配整个规则 r 的待分类样本的数目. 那么, 规则 r 的支持度为 $\mathrm{Sup}(r) = RN(r)/N$, 规则 r 的置信度为 $\mathrm{Con}(r) = RN(r)/TN(r)$, 规则 r 的覆盖度为 $\mathrm{Cov}(r) = RN(r)/CN(r)$, 所以, 规则 r 的适应度评价函数 $\mathrm{Fit}(r)$ 的定义如下:

$$\mathrm{Fit}(r) = \alpha \times \mathrm{Sup}(r) + \beta \times \mathrm{Con}(r) + \gamma \times \mathrm{Cov}(r) \quad (4)$$

其中, α, β 和 γ 分别表示支持度、置信度和覆盖度的权重.

3.3　多种群间的竞争机制

在多种群协同优化方法中, 竞争力强的种群获得计算资源的概率较大. 种群的竞争力主要取决于最佳适应度值和成长性两个因素. 在寻找最佳适应度值的过程中, 当一个种群的最佳适应度值较小时, 表明该种群处于进化过程的初始阶段, 距离最优解较远, 应将资源较多地分配给距离最优解较近的种群; 当一个种群再无成长性或成长性较差时, 如果该种群还没达到最优解, 表明它可能出现早熟现象, 趋于局部最优, 应将资源较多地分配给成长性较好的种群. 因此, 一个最佳适应度值大并且成长性好的种群对计算资源有更强的竞争力, 即获得计算资源的概率较大.

定义 5. 假设 N_p 表示普通种群的个数, $F^i_{\max}(k)$ 和 $F^i_{\max}(k-1)$ 分别表示第 i $(i = 1, 2, \cdots, N_p)$ 个种群当前和上一次进化过程的最佳适应度值, 那么, 第 i 个种群的成长性 G_i 的计算公式如下:

$$G_i = \frac{F^i_{\max}(k) - F^i_{\max}(k-1)}{F^i_{\max}(k-1)} \quad (5)$$

定义 6. 假设 $F^i_{\max}$ 表示第 i 个种群在整个进化过程中的最佳适应度值, 那么, 第 i 个种群的竞争力 C_i 的计算公式如下:

$$C_i = \alpha \frac{F^i_{\max}}{\sum\limits_{j=1}^{N_p} F^j_{\max}} + (1-\alpha) \frac{G_i}{\sum\limits_{j=1}^{N_p} G_j} \quad (6)$$

其中, $\alpha \in [0, 1]$ 为权重系数. α 的取值越小, 种群的

最佳适应度值对其竞争力的影响就越小, 而种群的成长性对其竞争力的影响就越大, 反之亦然.

3.4 多种群间的共享机制

优良种群由各个普通种群中适应度值较优的若干个体组成, 是一个优良种子库. 它由各个普通种群的进化过程共享. 多种群间的共享机制主要有两种方式: 良种迁移和良种交叉. 良种迁移是指各个普通种群在自身进化过程中, 直接从优良种群中引进若干优良种子个体替换本种群中的较劣个体. 良种交叉是指各个普通种群在自身进化过程中, 从优良种群中选取若干优良种子个体和本种群中的个体进行交叉繁殖, 用交叉所生成的较优个体替换本种群中用于本次交叉繁殖的个体.

4 实验

4.1 数据集

本文实验使用了如下 3 个标准文本数据集:

1) Reuters-21578 (Reuters). Reuters-21578 是文本分类中最常用的一个标准文本数据集[18], 其中, 包含 21 578 个文本, 共 135 个类别. 本文实验只选择那些至少属于一个类别且具有 "Lewis Split" 分割标记的文本, 这样的文本共有 16 895 个, 分别属于 106 个类别中的一个或者几个类别. 最后, 将这些文本的第一个类别标识作为文本的标准类别标识, 形成本文实验所采用的 Reuters 数据集.

2) 20 Newsgroups (20 NG). 20 Newsgroups 是一个常用的文本数据集[19], 收集了来自 20 个新闻组的 19 997 篇新闻. 该数据集有两个版本, 本文实验使用第二个版本. 第二个版本删除了第一个版本中重复文本和大部分文本头部, 包含 18 828 个文本.

3) 网页数据集 (Web). 网页数据集中的文本收集自 Google 的 Open Directory Project 项目[20]. 在本文实验中, 随机选取了其中的 35 个类别, 总共 5 035 个网页作为实验数据集. 所有的网页都使用 MSHTML Parser 进行处理以抽取其中的纯文本作为实验数据.

4.2 评价标准

本文实验从如下 4 个方面评价规则抽取方法:

1) 规则的数量: 是指采用某种方法对测试集进行分类规则抽取后, 最终得到分类规则的数目. 通常希望抽取到的分类规则越少越好.

2) 规则的准确率: 是指使用抽取到的分类规则对测试集进行分类, 最终得到分类的准确率. 通常希望抽取到的分类规则的准确率越高越好.

3) 规则的简易性: 是指抽取到的分类规则的平均长度. 通常希望分类规则越简单越好, 即平均长度越短越好.

4) 计算时间: 是指抽取分类规则所用的时间. 通常希望抽取分类规则所用的时间越少越好.

4.3 实验结果与分析

本文使用 CN2, FDT, Ant-Miner 和 LSVM 4 种方法与本文方法 (MPCO) 在上述 3 个标准文本数据集上进行比较实验. 本文实验是在处理器为奔腾 IV, 主频为 2.4 G, 内存为 512 M 的个人计算机上完成的. 本文实验采用 10 折交叉验证方法, 即将数据集随机划分成 10 个互不相交的子集, 进行 10 次分类规则抽取和测试, 依次轮流将其中一个子集作为测试集, 其他子集作为训练集, 取 10 次的平均值作为最终的实验结果. 具体实验结果见表 1.

表 1 5 种规则抽取方法的实验结果比较

Table 1 Comparison of experimental results of five rule extraction approaches

数据集	评价标准	CN2	FDT	Ant-Miner	LSVM	MPCO
Reuters	规则的数量	166.7	147.9	140.5	127.3	125.3
	规则的准确率	85.1 %	87.3 %	90.2 %	92.2 %	92.4 %
	规则的简易性	16.2	10.9	10.2	9.6	9.3
	计算时间 (秒)	1369.5	1212.8	1062.3	928.5	921.8
20 NG	规则的数量	132.6	105.2	99.6	95.6	94.3
	规则的准确率	81.5 %	84.1 %	86.9 %	90.3 %	90.6 %
	规则的简易性	12.3	9.9	9.5	8.9	8.6
	计算时间 (秒)	1208.5	1100.7	967.1	873.3	868.6
Web	规则的数量	75.3	62.9	58.8	55.1	53.2
	规则的准确率	81.8 %	84.3 %	87.1 %	90.5 %	90.9 %
	规则的简易性	9.5	7.8	7.3	7.0	6.8
	计算时间 (秒)	818.2	781.2	703.8	691.7	683.6

从表 1 中可以看出:

1) 在规则的数量方面

对于 Reuters 数据集而言, MPCO 和 LSVM 方法所抽取规则的数量明显少于 CN2、FDT 和 Ant-Miner 方法; MPCO 方法所抽取规则的数量略少于 LSVM 方法. 对于 20 NG 数据集而言, CN2 方法所抽取规则的数量明显多于其他 4 种方法; MPCO 和 LSVM 方法所抽取规则的数量接近. 对于 Web 数据集而言, MPCO 方法所抽取规则的数量少于其他 4 种方法.

2) 在规则的准确率方面

对于 Reuters 数据集而言, MPCO 方法所抽取规则的准确率明显高于 CN2、FDT 和 Ant-Miner 方法, 略高于 LSVM 方法. 对于 20 NG 和 Web 数据集而言, MPCO 和 LSVM 方法所抽取规则的准确率明显高于其他 3 种方法; MPCO 方法所抽取规则的准确率略高于 LSVM 方法.

3) 在规则的简易性方面

对于 3 个数据集而言, MPCO 和 LSVM 方法所抽取规则的平均长度都比其他 3 种方法短, MPCO 方法所抽取规则的平均长度比 LSVM 方法略短; 尤其是对于 Reuters 和 20 NG 数据集而言, MPCO 方法所抽取规则的平均长度明显比 CN2 方法短.

4) 在计算时间方面

对于 3 个数据集而言, MPCO 方法所用的计算时间少于其他 4 种方法; 特别是对于 Reuters 和 20 NG 数据集而言, MPCO 方法所用的计算时间明显少于 CN2 和 FDT 方法.

我们分析本文提出的 MPCO 方法效率高的原因主要有以下两点:

1) 基于信息熵的种群初始化方法

该方法基于信息熵理论构建而成, 根据特征的平均信息量确定将特征插入到当前规则中的概率. 该方法能有效减少规则抽取所用的时间.

2) 基于多种群协同优化的规则抽取方法

该方法是借鉴自然界中各种生物之间通过信息交换来取长补短并共同进化的现象设计出来的. 多种群协同优化系统由计算环境中的多个普通种群和一个优良种群构成, 各个普通种群通过相互竞争机制在计算环境中竞争计算资源. 同时, 各个普通种群将进化得到的优良个体贡献出来, 组成优良种群. 普通种群可以从优良种群中获取优良个体, 以改善本种群的品质. 这样, 多种群优化方法通过种群之间的相互竞争和良种共享机制提高优化方法的效率.

5 结论和展望

本文首先简要介绍了文本分类, 并给出了基于规则抽取的 Web 文本分类的流程. 其次, 针对该流程中的文本分类规则抽取, 本文提出了一种基于多种群协同优化的文本分类规则抽取方法, 并给出了该方法的优化框架. 然后, 详细介绍了基于信息熵的种群初始化, 给出了多种群协同优化系统的结构和多种群协同优化方法的流程, 并详细介绍了适应度评价函数, 分别介绍了多种群间的竞争和共享机制. 最后, 使用 CN2、FDT、Ant-Miner 和 LSVM 4 种方法与本文提出的方法在 Reuters, 20 Newsgroups 和 Web 3 个标准文本数据集上进行了比较实验, 给出了评价标准, 对实验结果做了详细的比较分析. 实验结果表明, 本文提出的文本分类规则抽取方法所抽取规则的数量少, 准确率高, 平均长度短; 同时, 本文方法所用的计算时间少, 抽取分类规则的速度快, 适用于大规模数据集.

以后的研究工作:

1) 本文方法中参数的设定: 更加合理地设定本文方法中的各个参数, 使该方法能更加高效地完成优化问题的求解.

2) 多种群协同优化方法中竞争和共享机制的改进: 通过竞争和共享机制的改进进一步提高本文方法的优化效率.

3) 分类规则的简化: 对抽取到的规则进行简化, 使分类规则的平均长度进一步缩短, 从而提高分类的效率.

References

1 Shin K, Abraham A, Han S Y. Improving *k*NN text categorization by removing outliers from training set. In: Proceedings of the 7th International Conference of Computational Linguistics and Intelligent Text Processing. Mexico City, Mexico: Springer, 2006. 563–566

2 Kim S B, Han K S, Rim H C, Myaeng S H. Some effective techniques for naive Bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, 2006, **18**(11): 1457−1466

3 Martens D, Huysmans J, Setiono R, Vanthienen J, Baesens B. Rule extraction from support vector machines: an overview of issues and application in credit scoring. *Studies in Computational Intelligence*, 2008, **80**: 33−63

4 Carter P H. The Rocchio classifier and second generation wavelets. In: Proceedings of the International Society for Optical Engineering. Orlando, USA: SPIE, 2007. 1−11

5 Esuli A, Fagni T, Sebastiani F. Boosting multi-label hierarchical text categorization. *Information Retrieval*, 2008, **11**(4): 287−313

6 Wang J, Yao Y, Liu Z J. Web page automatic categorization based on non-linear SVM decision tree. *Journal of Computational Information Systems*, 2008, **4**(2): 449−454

7 Qiu J T, Tang C J, Zeng T, Qiao S J, Zuo J, Chen P. A novel text classification approach based on enhanced association rule. In: Proceedings of the 3rd International Conference on Advanced Data Mining and Applications. Harbin, China: Springer, 2007. 252−263

8 Goyal R D. Knowledge based neural network for text classification. In: Proceedings of IEEE International Conference on Granular Computing. San Jose, USA: IEEE, 2007. 542−547

9 Fujino A, Ueda N, Saito K. Semisupervised learning for a hybrid generative/discriminative classifier based on the maximum entropy principle. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(3): 424−437

10 Yin S Q, Wang F, Xie Z, Qiu Y H. Study on web-page classification algorithm based on rough set theory. In: Proceedings of International Symposium on Information Processing/International Pacific Workshop on Web Mining and Web-based Application. Moscow, Russia: IEEE, 2008. 202−206

11 Wang Jian-Hui, Wang Hong-Wei, Shen Zhan, Hu Yun-Fa. A simple and efficient algorithm to classify a large scale of texts. *Journal of Computer Research and Development*, 2005, **42**(1): 85−93
(王建会, 王洪伟, 申展, 胡运发. 一种实用高效的文本分类算法. 计算机研究与发展, 2005, **42**(1): 85−93)

12 Wang Yu, Wang Zheng-Ou. Text categorization rule extraction based on fuzzy decision tree. *Journal of Computer Applications*, 2005, **25**(7): 1634−1637
(王煜, 王正欧. 基于模糊决策树的文本分类规则抽取. 计算机应用, 2005, **25**(7): 1634−1637)

13 Clapk P, Boswell R. Rule induction with CN2: some recent improvements. In: Proceedings of the 6th European Working Session on Learning. Porto, Portugal: Springer, 1991. 151−163

14 Parpinelli R S, Lopes H S, Freitas A A. A data mining with an ant colony optimization algorithm. *IEEE Transactions on Evolutionary Computing*, 2002, **6**(4): 321−332

15 Fung G, Sandilya S, Rao R B. Rule extraction from linear support vector machines. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA: ACM, 2006. 32−40

16 de la Cruz G J P. An unsupervised learning rule for class discrimination in a recurrent neural network. In: Proceedings of the 16th International Conference on Artificial Neural Networks. Athens, Greece: Springer, 2006. 415−424

17 Tang Hua, Zeng Bi-Qing. Research on method of text classification rule extraction based on genetic algorithm and entropy. *Acta Scientiarum Naturalium Universitatis Sunyatseni*, 2007, **46**(5): 18−21
(唐华, 曾碧卿. 基于遗传算法和信息熵的文本分类规则抽取方法研究. 中山大学学报 (自然科学版), 2007, **46**(5): 18−21)

18 Bekkerman R, EI-Yaniv R, Tishby N, Winter Y. On feature distributional clustering for text categorization. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New Orleans, USA: ACM, 2001. 146−153

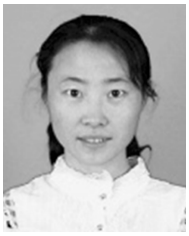
19 Zelikovitz S, Hirsh H. Using LSI for text classification in the presence of background text. In: Proceedings of the 10th International Conference on Information and Knowledge Management. Atlanta, USA: ACM, 2001. 113−118

20 Zeng H J, Chen Z, Ma W Y. A unified framework for clustering heterogeneous web objects. In: Proceedings of the 3rd International Conference on Web Information Systems Engineering. Singapore, Singapore: IEEE, 2002. 161−170

刘 赫 2005 年获得吉林大学计算机科学与技术学院硕士学位. 2009 年获得吉林大学计算机科学与技术学院博士学位. 主要研究方向为数据挖掘和文本挖掘.
E-mail: liuhe1980@163.com
(**LIU He** He received his master and Ph. D. degrees from the College of Computer Science and Technology, Jilin University in 2005 and 2009. His research interest covers data mining and text mining.)

刘大有 吉林大学计算机科学与技术学院教授. 主要研究方向为知识工程、专家系统与不确定性推理、时空推理、分布式人工智能、多 Agent 和移动 Agent 系统、数据挖掘与多关系数据挖掘、数据结构与计算机算法. 本文通信作者.
E-mail: dyliu@jlu.edu.cn
(**LIU Da-You** Professor at the College of Computer Science and Technology, Jilin University. His research interest covers knowledge engineering, expert system and uncertainty reasoning, spatio-temporal reasoning, distributed artificial intelligence, multi-agent systems and mobile agent systems, data mining and multi-relational data mining, data structures, and computer algorithms. Corresponding author of this paper.)

裴志利 内蒙古民族大学计算机科学与技术学院博士, 副教授. 2008 年获得吉林大学计算机科学与技术学院博士学位. 主要研究方向为生物信息学和文本挖掘.
E-mail: zhilipei@sina.com
(**PEI Zhi-Li** Ph. D. and associate professor at the College of Computer Science and Technology, Inner Mongolia University for Nationalities. He received his Ph. D. degree from Jilin University in 2008. His research interest covers bioinformatics and text mining.)



高 滢 吉林大学计算机科学与技术学院博士, 讲师. 2008 年获得吉林大学计算机科学与技术学院博士学位. 主要研究方向为数据挖掘和统计关系学习.
E-mail: gying@jlu.edu.cn
(**GAO Ying** Ph. D. and lecturer at the College of Computer Science and Technology, Jilin University. She received her Ph. D. degree from Jilin University in 2008. Her research interest covers data mining and statistical relational learning.)