一种基于云模型的特征选择参数优化研究

宋 丽,张震雷,杨新凯
(上海师范大学 信息与机电工程学院,上海 200234)

摘 要: 常用特征选择方法面临着特征子集空间大小难以确定的问题,取不同的 k 值,它们的分类效果是相差很大的。粒子群优化算法存在收敛快、获得的是局部最优值而不是全局最优值的问题。针对上述问题,结合云模型的理论知识,提出一种基于云模型的特征选择方法。该算法的适应度函数是通过精确率这一评价指标计算的,将权重分为三个类别来动态确定惯性权重。采用模糊期望交叉熵对原始的特征子集空间进行预选,将预选后的特征子集作为原始特征空间采用改进的特征选择方法,根据模糊期望交叉熵的大小来初始化粒子的种群数及采用迭代变化的阈值作为控制算法的结束条件。实验结果证明了该方法的有效性和可行性。

关键词: 特征选择; 参数优化; 粒子群; 云模型
中图分类号: TP391 **文献标识码:** A **文章编号:** 1673-629X(2019)03-0093-04
doi: 10.3969/j.issn.1673-629X.2019.03.020

Research on Optimization of Text Feature Selection Parameters Based on Cloud Model

SONG Li,ZHANG Zhen-lei,YANG Xin-kai
(School of Information and Electromechanical Engineering,Shanghai Normal University,Shanghai 200234,China)

Abstract: Common feature selection methods are faced with the problem that the size of feature subset space is difficult to determine. Taking different k values,their classification effects are quite different. PSO algorithm has the problem of fast convergence and obtaining the local optimal value instead of the global optimal value. To solve the above problems,in combination of the theoretical knowledge of cloud models,we propose a feature selection method based on cloud model. The fitness function is calculated by the accuracy rate evaluation index. The weight is divided into three categories to dynamically determine the inertia weight. In this paper,the original feature subset space is preselected using the fuzzy expectation cross entropy,and the pre-selected feature subset is used as the original feature space to adopt an improved feature selection method. According to the size of the fuzzy expected cross entropy,the population of particles is initialized. The iteratively changing threshold serves as an end condition for the control algorithm. The experiment shows that the proposed method is feasible and effective.

Key words: feature selection; parameter optimization; particle swarm; cloud model

0 引 言

随着互联网技术的蓬勃发展,网络中的文本正以指数级的速度增长。因此,提取大数据量下的有效信息,提高文本分类的高效性和准确性、提高高质量和智能化的文本分类、满足用户所需要的信息服务具有重要的意义。特征选择和特征降维是实现特征研究的关键步骤。文中主要研究的是特征选择技术。文本特征选择技术[1]是实现文本分类的关键问题之一[2]。常用的方法[3]是降低短文本特征维数和去除冗余无关的特征[4]。文中通过常用的文本特征选择方法[5]进行预选;提出了一种基于云模型[6]优化的特征选择方法,预选的特征子集作为第二次特征选择的原始特征空间,解决了特征子集规模大小难以确定的问题。

1 研究现状

最原始的文本分类依赖于人工,该分类方式对参与文本分类工作人员的要求较高,需要他们对各领域的知识都有良好的认知和掌握。该分类方法效率极

收稿日期: 2018-04-15 **修回日期:** 2018-08-16 **网络出版时间:** 2018-12-19
基金项目: 国家自然科学基金(61572326)
作者简介: 宋 丽(1991-),女,硕士研究生,CCF会员(90940G),研究方向为文本分类;杨新凯,研究生导师,研究方向为自然语言处理、普适计算等。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181219.1532.058.html

低、成本较高，不适用于大规模文本分类的场景。20 世纪 50 年代末，相关学者开始对文本分类方法进行研究。1961 年，Maron 等[7]发表了一篇关于文本分类的具有里程碑意义的文章，推动了文本分类的进展。20 世纪 70 年代初，Rocchio[8]将线性分类器应用于文本分类领域，该方法通过用户的反馈信息不断地修正分类器的权重参数，来提高线性分类器的学习与分类效果。90 年代末，相关研究学者[9]将机器学习方法应用于文本分类任务中。20 世纪末，Yoav 和 Robert[10]提出了提升树分类算法，进而形成强分类器，这一算法的提出将文本分类算法推向了高峰。国内关于文本分类方法的研究起步较晚。目前，研究学者分别从特征选择、权重、分类模型等角度对其进行研究。对特征选择的研究，主要集中于特征扩展、特征选择、特征权重计算和分类算法设计这几个方面。常见的特征选择方法大都是基于常用方法，包括模式匹配法、词频法、卡方统计法、互信息法等。例如，姚旭等[11]分析了特征选择方法的框架，从搜索策略和评价准则两个角度[12]对特征选择方法进行分析和总结，分析出对特征选择的影响因素，并指出了实际应用中需解决的问题；徐刚等[13]提出一种根据速度信息自适应调整参数的粒子群优化算法，该算法可解决复杂非线性优化时搜索失败的问题。

以上研究主要依赖于传统的特征选择、特征提取及特征技术的组合使用，云模型的应用相对较少。文中主要通过传统特征选择技术进行预选，提出一种基于云模型的粒子群优化方法进行二次选择。

2 参数优化模型

2.1 粒子编码

粒子编码一般有 UTF-8 编码、二进制编码、GBK、实数编码等形式。粒子群优化算法离不开编码问题的设置。其中，二进制是最常用的编码方式，二进制编码具有易实现，编、解码简单操作等优势，得到了广泛应用。文中也采用二进制编码，用 0、1 表示粒子的位置信息，1 表示该粒子被选中，0 表示该特征被丢弃；用特征模糊熵值初始化粒子群的初始化种群，如：含 200 个特征的文本集，5 个种群数，生成一个初始矩阵 [5 * 200]，矩阵形式如下所示。

$$\begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & \cdots & 199 & 200 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & \cdots & 1 & 0 \\ 1 & 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 1 & 0 & 1 & \cdots & 1 & 1 \\ 1 & 1 & 1 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \end{array}$$

2.2 粒子群技术

在特征选择的过程中，特征子集空间规模的确定成为一个难点。一般采用最优化的方法，对特征集空间进行全局寻优。文中采用粒子群算法(PSO)。PSO 是一种全局优化方法，将优化问题的每个解作为搜索空间的一个“粒子”，每个粒子由优化函数决定的适应度来评价粒子当前位置的优劣。每个粒子有一个速度来决定各自移动的方向和距离。在每次迭代中，粒子跟踪两个“极值”来更新自己：一个是个体极值(pBest)，一个是全局极值(gBest)。

第 i 个粒子找到上述两个极值，更新自己的速度和位置：

$$v_{id}^{t+1} = w\, v_{id}^{t} + c_1\, r_1 (P_{id} - x_{id}^{t}) + c_2\, r_2 (P_{gd} - x_{id}^{t}) \tag{1}$$

$$x_{id}^{t+1} = x_{id}^{t} + v_{id}^{t+1} \tag{2}$$

其中，w 为惯性权重；v_{id}^{t} 为第 t 次搜索中粒子 i 第 d 维的飞行速度；x_{id}^{t} 为第 t 次搜索中粒子 i 第 d 维的位置信息；P_{id} 为当前粒子 i 在 d 维上找到最好的位置；P_{gd} 为搜索结束后所有粒子在第 d 维上找到的最好位置；c_1，c_2 为受外界因素影响的学习因子，一般设为相同体，$c_1 = c_2 = 2$；r_1，r_2 为 [0，1] 区间上的随机数。

惯性权重：

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{\mathrm{iter}_{\max}} * t \tag{3}$$

其中，$w_{\max}$ 为最大惯性权重；$w_{\min}$ 为最小惯性权重；$\mathrm{iter}_{\max}$ 为最大迭代次数；t 为当前迭代次数。

2.3 基于云模型的粒子群方法

针对粒子群算法[9]存在早熟和不收敛的问题，通过权重生成策略来改变惯性权重的方法，保证收敛速度与全局收敛之间的折中。

2.3.1 适应度函数

特征选择主要是剔除不相关或无用的特征来对文本进行预处理。文本分类的三个评价指标中，准确率最能衡量特征与类别的关系程度。因此这里的适应度函数用准确率这一评价指标，公式如下：

$$f = \sum_{i=1}^{m} \frac{n_i}{N} \tag{4}$$

其中，m 为总类数；N 为总的文本数；n_i 为类别 i 被正确分类的文本数。

2.3.2 权重生成策略

结合云模型的优点，提出一种基于云模型的粒子群优化算法，以动态确定惯性权重来避免粒子群的早熟收敛问题。

给出如下定义：f_i 为粒子 x_i 的适应度函数值，$f_{\mathrm{avg}} = \frac{1}{N}\sum_{i=1}^{N} f_i$（$N$ 为粒子群的种群数）为粒子群的平均适应度值，f_{pg} 为当前全局最优值的函数值。根据粒子的适应度函数值将粒子群分为三种情况进行研究分析，每

种情况赋予不同的惯性权重,惯性权重 w 的生成方法为:

(1) $f_i > f_{pg}$: 当粒子的适应度值大于当前全局最优值时,表示该粒子在群中是最好的,解与全局最优值很接近。为了加快全局收敛速度,缩短运行时间,这种情况下使惯性权重最小。此时惯性权重为:

$$w = w_{min} \tag{5}$$

(2) $f_{avg} < f_i < f_{pg}$: 当粒子的适应度值大于粒子群的平均适应度值且小于当前全局最优值时,说明这些粒子是普通粒子。采用云模型理论,根据云模型条件的云发生器的非线性动态特性来改变粒子的惯性权重。惯性权重的确定公式为:

$$\begin{aligned} E_x &= f_{pg} \\ E_n &= | f_i - f_{avg} | / c_1 \\ H_e &= E_n / c_2 \\ E_n' &= \mathrm{normrnd}(E_n, H_e) \\ w &= w_{max} - w_{min} * \exp\left(\frac{-(f_i - E_x)^2}{2(E_n')^2}\right) \end{aligned} \tag{6}$$

从公式中可以得到,w 与适应度函数值成反比,实现了较优的粒子与较小的 w 值的对应。经实验验证,式中 $c_1 = c_2 = 4$。

(3) $f_i < f_{avg}$: 当粒子的适应度值小于粒子群的平均适应度值时,表明粒子在群体中是最差的。若要进行全局寻优,建议采用较大的惯性权重。惯性权重的确定公式为:

$$w = w_{max} \tag{7}$$

通过阅读文献资料,总结出惯性权重通常取 [0.4,0.95] 这个区间时,粒子的优化性能飞速提高。因此,$w_{min} = 0.4$, $w_{max} = 0.95$。

2.3.3　基于云模型的粒子群技术

基于云模型的特征选择方法由两阶段完成。第一阶段采用 AFECE 对原始的特征集进行预选; 第二阶段将第一阶段预选出的特征子集作为第二阶段的原始集,采用云模型特征选择方法来进行第二次特征选择。

算法描述如下:

输入: 原始特征集 X'

输出: 特征子集

(1) 特征预选

(2) 初始化粒子群各参数

(3) 计算惯性权重,更新粒子的速度和位置

(4) if$(f_i^{t+1} > f_{P_i})$: $P_i = x_i^{t+1}$

else : $f_{P_i} = f_i^{t+1}$

(5) if$(f_i^{t+1} < f_{P_g})$: $P_g = P_i$

else : $f_{P_g} = f_{P_i}$

(6) 根据结束条件来判断是否满足。否则返回到(3)

算法的结束条件是根据在全局寻优的过程中,不断变化的阈值来判断的。图 1 为算法的流程框图。

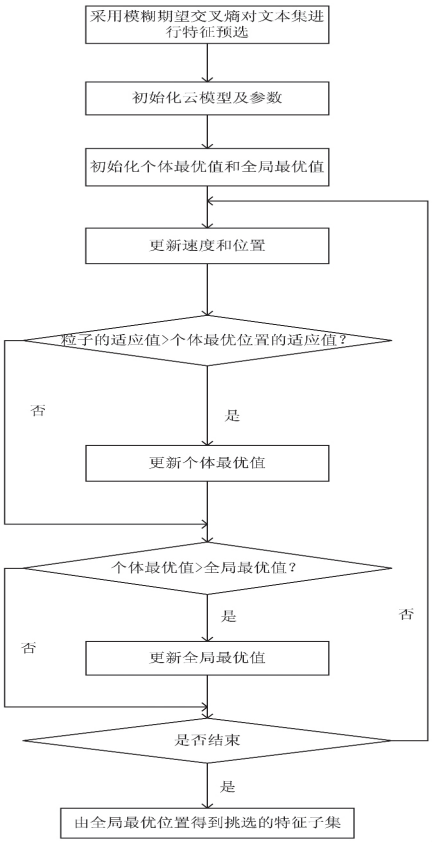

图 1　CPSO 特征选择方法

3　实验与结果分析

利用从 http://kdd.ics.uci.edu 网站上下载的 reuters-21578 数据集进行仿真,训练集占 80%,测试集占 20%。

3.1　预选的特征子集性能

reuters-21578 数据集,经 MI、CHI、ECE、AFECE 四种特征选择方法进行特征预选,效果如图 2 和图 3 所示。

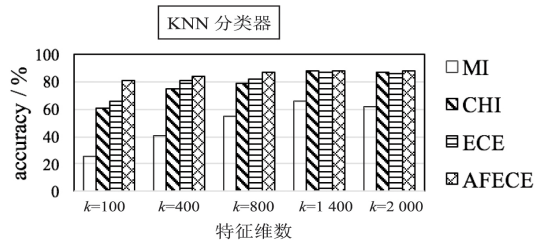

图 2　不同特征选择方法在不同特征维数下的精确率

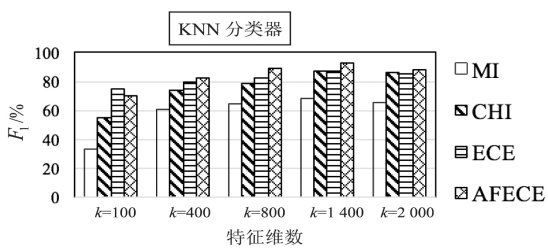

图 3　不同特征选择方法在不同特征维数下的 F_1 值

3.2 基于云模型的粒子群优化方法

在预选特征子集的基础上进行二次特征选择，初始化各参数如表 1 所示。

表 1 初始化各参数

参数	数值
k（特征子集大小）	400
N（种群数）	10
$\text{iter}_{\max}$（最大迭代次数）	40
c_1，c_2（影响因子）	4
$v_{\max}$（最大飞行的速度）	4
$v_{\min}$（最小飞行速度）	-4
T_{th}（迭代变化阈值）	4
v_{id}^0（粒子的初始速度）	0

表 1 是二次特征进行选择时各参数大小的设置。第一阶段的特征子集空间选择 400 作为第二阶段的原始特征空间。这里的种群数与类别数是一一对应的。其他各参数是通过数据训练的调试及经验进行设置的。表 2 是第二阶段的三种特征选择方法与第一阶段的 AFECE 特征选择方法所得的实验结果。可以得出，SPSO 所需的特征维数最少，就能达到较好的分类效果，但它的性能是四种方法里最低的。AFECE 特征选择方法性能优于 PSO、SPSO，但它是以牺牲空间换来的。CPSO 特征选择方法性能有所改善，在精确率方面提高了 4.4%，且所需的特征空间小。

表 2 reuters-21578 语料上各参数优化方法的对比实验结果

方法	P /%	R /%	F_1 /%	特征维数(int)
CPSO	88.34	82.34	85.27	254
SPSO	77.15	74.24	76.64	163
PSO	80.85	73.87	78.56	209
AFECE	83.94	79.89	82.94	400

4 结束语

基于特征子集空间规模大小难以确定的问题，提出了一种基于云模型的特征选择算法，由以下 2 个阶段组成：第一阶段比较 MI、CHI、ECE、AFECE 四种特征选择方法在性能上的优缺点，得到 AFECE 特征选择方法性能最好；第二阶段通过 AFECE 特征选择方法选出的特征子集作为这一阶段的原始特征空间，提出基于云模型的粒子群算法进行二次特征选取，避免了粒子过早收敛达到局部最优而不是全局最优的不足。实验结果证明了该方法的有效性和可行性。文中在特征选择方面做了一定程度的研究，在分类器设计方面没有进行尝试。因此提出适合的分类方法和改进分类器参数的确定将是今后研究的工作之一。除此之外，将该方法应用于产品智能推荐研究领域也是下一步重点研究的工作。

参考文献：

[1] WANG Bingkun，HUANG Yongfeng，YANG Wanxia，et al. Short text classification based on strong feature thesaurus [J]. Journal of Zhejiang University：Science C，2012，13(9)：649-659.

[2] MARCHEGGIANI D，TACKSTROM O，ESULI A，et al. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining [M] //Advances in information retrieval. [s. l.]：Springer International Publishing，2014：273-285.

[3] KIM S B，HAN K S，RIM H C，et al. Some effective techniques for naive Bayes text classification [J]. IEEE Transactions on Knowledge and Data Engineering，2006，18(11)：1457-1466.

[4] WU Mingfen，HAN Haohan，SI Yanfei. Properties and axiomatization of fuzzy rough sets based on fuzzy covering [C] //International conference on machine learning and cybernetics. Xi'an，China：IEEE，2012：184-189.

[5] PANG Bo，LEE L，VAITHYANATHAN S. Thumbs up?：sentiment classification using machine learning techniques [C] //Proceedings of the conference on empirical methods in natural language processing. Philadelphia：ACL，2002：79-86.

[6] LI Deyi，LIU Changyu，LIU Luying. Study on the universality of the normal cloud model [J]. Engineering Sciences，2005，3(2)：18-24.

[7] MARON M E，KUHNS J L. On relevance，probabilistic indexing and information retrieval [J]. Journal of the ACM，1960，7(3)：216-244.

[8] ROCCHIO J. Relevance feedback in information retrieval [C] //The smart retrieval system：experiments in automatic document processing. New Jersey：Prentice-Hall，1971：313-323.

[9] 赵小华. KNN 文本分类中特征词权重算法的研究 [D]. 太原：太原理工大学，2010.

[10] FREUND Y，SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences，1997，55(1)：119-139.

[11] HU Y，YANG C，JI C，et al. Efficient snapshot KNN join processing for large data using MapReduce [C] //2016 IEEE 22nd international conference on parallel and distributed systems. Wuhan：IEEE，2016：713-720.

[12] 姚 旭，王晓丹，张玉玺，等. 特征选择方法综述 [J]. 控制与决策，2017，27(2)：161-166.

[13] 徐 刚，瞿金平，杨智韬. 一种改进的自适应粒子群优化算法 [J]. 华南理工大学学报：自然科学版，2015，36(9)：6-10.