# PSO-GP 中文文本情感分类方法研究

黄 熠[1] 王 娟[2]
(解放军特种作战学院 广州 510502)[1] (国网信息通信产业集团有限公司 北京 102211)[2]

**摘 要** 中文文本的情感倾向分析是网络舆情信息挖掘和分析的关键技术之一。提出了一种粒子群-高斯过程算法(PSO-GP)的中文文本情感倾向分类方法,采用粒子群优化算法(Particle Swarm optimization,PSO)进行高斯过程(Gaussian Process)超参数的最优搜索,解决了传统高斯过程中共轭梯度法迭代次数难确定、对初值依赖性强和易陷入局部极小值等问题。首先采用多线程网络爬虫技术采集文本数据组成语料库,构建特定领域情感词典,然后通过情感词匹配选择最有效的特征,降低数据维度,并利用 TF-IDF 算法计算特征词的权重以生成特征向量。最终,将测试样本输入 PSO-GP 分类模型。实验结果表明,与传统 GP 方法相比,提出的改进高斯过程分类模型的分类准确率提高了近 15%。

**关键词** 中文文本情感分类,网络爬虫,情感词典,粒子群优化算法,高斯过程
**中图法分类号** TP39 **文献标识码** A

## Research on Chinese Texts Sentiment Classification Approach Based on PSO-GP

HUANG Yi[1] WANG Juan[2]
(Chinese People's Liberation Army Special Warfare School,Guangzhou 510502,China)[1]
(State Grid Information Telecommunication Group Co.,Ltd,Beijing 102211,China)[2]

**Abstract** The Chinese texts sentiment orientation analysis is one of the key technologies for network public opinion information mining and analysis. This paper proposed a Chinese texts sentiment classification method based on particle swarm optimization-Gaussian process (PSO-GP) algorithm, which employs optimal hyper parameter search. It solves problems of traditional Gauss iteration process, like difficultly to determine conjugated gradient, strongly dependent on initial value and easily to fall into local minimum. It can collect data set text corpus to construct a domain specific emotion dictionary using multi-threaded web crawler technology, select the most effective features by emotional words, reduce the data dimension, and generate feature vector from feature words by TF-IDF algorithm. Experimental results show that the classification accuracy of the improved classification model is improved by nearly 15%.

**Keywords** Chinese texts sentiment classification, Web crawlers, Semantic lexicon, Particle swarm optimization, Gaussian process

## 1 引言

随着 20 世纪互联网的诞生,互联网中涌现了大量针对产品、热点事件以及焦点人物的主观性信息,其表现形式大多为非结构化或者半结构化的评论文本形式[1]。网络上的信息越来越受到企业和政府的关注。国家政府管理部门根据网络上广大网民发布的言论,可以更好地了解群众意愿,从而更好地服务大众并制定更加人性化的政策;企业则可以根据产品的相关反馈和评论了解大众对产品以及服务的意见和建议,以此改变售后服务和产品质量[2]。如何处理这些消息来获得想要的知识,在近年来得到了广泛的关注。

文本情感倾向分类的主要目的是让计算机自动识别主观性文本的情感倾向性及其强度,分类结果分为正面和负面两种[3]。国外对英文情感分类的研究起步较早,并取得了很大的进展。而国内对于中文文本情感分类的研究较晚,不过也取得了一些相关的成果。

在中文文本的情感倾向分类研究中,主要有基于语义统计和基于机器学习两种方法[4]。基于语义统计的方法主要是通过统计文本中情感词的词汇倾向值进行文本分类[5];利用机器学习方法进行情感倾向分类是目前的一个热门研究方向,主要的方法有 KNN、神经网络、朴素贝叶斯、支持向量机等[6-7]。利用机器学习进行文本倾向性分类大体可归纳为 2 个步骤:1)文本特征选择和特征向量生成;2)分类器模型构建以及对语料进行训练和分类,其中不同的特征选择方法、不同的权重计算以及不同的机器学习方法会对分类结果造成不同程度的影响[7]。

上述研究虽然在中文文本情感倾向分类方面取得了一定的效果,但仍然存在特征提取不适当、检测率和检测精度不高、算法复杂度高等问题。针对上述问题,本文引入粒子群算法(PSO)[8],通过优化高斯过程的超参数形成粒子群-高斯过程分类算法(PSO-GP),并以实际知名电商网站商品的评价为语料库进行特征提取和权重计算,并构造特征向量,将其作为

黄 熠(1987-),女,助理工程师,主要研究方向为网络安全、数据挖掘;王 娟(1984-),女,硕士生,工程师,主要研究方向为自然语言处理,E-mail:46628376@qq.com(通信作者)。

实验数据应用到所提出的优化分类模型中。该方法提高了中文文本情感倾向问题的分类精度，在实际问题的处理上具有良好的应用前景。

## 2　基于网络爬虫和 TF-IDF 的中文文本特征提取

中文文本情感对事物的倾向是两面性的，即正面和负面、褒义和贬义等。由此可见，中文文本情感倾向分类问题是一个二分类问题。要构建一个性能良好的文本情感分类模型，需要做好几方面工作：首先，选取可靠的语料来源，得到明显的情感倾向；其次，找到情感词的本质特征，良好地表示特征数据；最后，构造一个性能良好的分类器模型，将正面和负面情感词进行高精度的分类。

### 2.1　基于网络爬虫的语料采集

随着网络爬虫技术[9]以及 DOM 规范[10]的不断发展，自动采集数据也变得相对简便。

为了能够高效获取中文文本语料，本文采用网络爬虫技术对文本语料进行采集。以知名电商的空调产品评论为例，将网页的源 html 文本渲染成 DOM(Document Object Model)节点树，获取包含目标内容的 URL，得到 html 文本，取出所需要的内容，包括评论文本、评论日期、评论得分以及相应的超链接等，如图 1 所示。具体的实现步骤如下：

1)使用 HttpClient 技术首先获取淘宝、京东、苏宁和国美 4 个网站的“空调”搜索结果页面文本，即 html 文本。

2)使用 jsoup 解析第 1)步获得的 html 文本，并得到 DOM 节点树，然后获取评论的 URL。

3)根据第 2)步得到的评论 URL 得到评论页面的 html 文本。

4)解析上一步得到的 html 文本，获取评论的得分、评论内容和评论日期，并将其写入数据库，考虑到评论的有效性，这里只筛选字数多于 50 的评论。

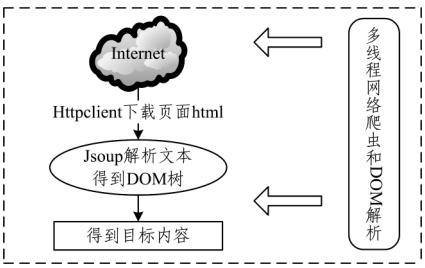



图 1　爬虫自动采集语料流程图

本文不再赘述目前网络爬虫技术和 DOM 规范。以 jd/tb/gm/sn 分别代表电商京东、淘宝、国美和苏宁，score 为用户对空调的打分程度，emotion 为文本情感倾向，取值 0 或 1(+1 或−1)。通过 Myeclipse 运行爬虫程序，得到的部分数据如图 2 所示。

| id | source | content | score | emotion |
|---|---|---|---|---|
| 1022 | tb | 夏天制冷效果不错，外观也还可以，本来觉得还不错，但那不久天见才开了下 | 1 | 1 |
| 1023 | jd | 我现在使用的就是1999年结婚时买的美克斯空调！十五年来从未修过，连氟都 | 5 | 0 |
| 1024 | tb | 已经安装好了，很麻利的安装师傅，半个小时就了2个空调，试用了，热风很快 | 5 | 0 |
| 1025 | gm | 送货挺快的，安装师傅态度也不错，提前的了时间。空调本身没什么问题，安 | 5 | 0 |
| 1026 | jd | 太坑人了，一直都在京东买东西，结果这次太让分了。买台空调才1499元，算 | 1 | 1 |
| 1027 | tb | 外机与内机 个人感觉 噪音有点大 每个人的承受感觉不同 暂时只适用了制热 | 5 | 0 |
| 1028 | sn | 首先你得知道便宜没好货 其次你得知道美克斯这种不大不小的企业的毒眼一般 | 1 | 1 |
| 1029 | gm | 空调挺好看的，快递人员搬得的安装人员不如厂家的负责，竟然让用马车这么 | 5 | 0 |
| 1031 | gm | 选对是好东西，就是安装师傅调度室外机有点大，两个做不下得到抽着才行！其 | 5 | 0 |
| 1032 | jd | 双十一给爸爸买的，客服态度很好，很快就排了安装。效果听说还可以，制热 | 5 | 0 |
| 1033 | tb | 发货速度快，尽快在双十一购买的，也就两三天到货。售后服务很好，主动电 | 5 | 0 |
| 1034 | sn | 我家是第二次购买了，还有别人买过一台，三部了。想告诉大伙：做人要实事 | 5 | 0 |
| 1036 | tb | 第一次买美克斯空调，没想到调到28度后空调一晚不停，被冻醒后发现温度已 | 1 | 1 |
| 1038 | gm | 京东送货很好，可惜美克斯不咋的，1.安装的时候给我开了个超级大洞。2.挂 | 1 | 1 |
| 1040 | jd | 空调真心挺好，物流快，运边还没到质守候已经提前打电话过来安装了，真心 | 5 | 0 |

图 2　部分数据库语料

### 2.2　中文文本预处理及特征选择

中文不像英文那样每个词汇之间被空格分开，所以要先进行分词处理。本文采用了最大匹配算法对中文文本进行分词，该方法属于基于字符串匹配的分词方法，需要分词词典的支持。本文采用了国家语言文字工作委员会发布的《现代汉语常用词表(草案)》(LCWCC)作为分词词典，该词典搜集了现在日常生活中使用频率较高的 56008 个词汇，能够满足分词的需要。

在特征选择方法上，本文采用了文献[12]中的基础情感词(BSL)作为本文选用的情感词典，该情感词共有 5281 个，其中褒义词 2807 个，贬义词 2474 个。这个词典是在 How-Net 情感词语集[13]的基础上，利用其提供的义原计算两个词的相似度，根据词与正向和负向种子词的平均相似度的差来判定词的情感倾向得到的一个情感词典。

就文本情感分类来说，特征选择就是选择出一些最有效的特征来降低数据集维度的过程。本文将上述构造的情感词典与文本中的情感词进行匹配，并作为最终的文本特征。情感文本处理的流程如图 3 所示。该方法操作简单，对后续构建优化的高斯过程分类器以及提高训练分类效果具有重要意义。

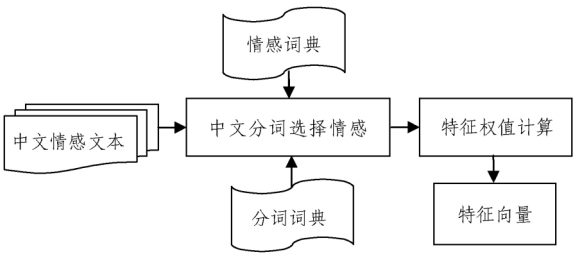



图 3　情感文本处理的流程

### 2.3　特征权值计算及表示

特征权重表示特征词在分类中所占的分量，是文本分类过程中的一个重要依据。常用的权重计算方法有布尔权重、绝对词频权重、归一化词频权重计算方法[14]。以上 3 种权重计算方式只是对文本的信息进行初步统计，并没有考虑每个特征集在整个文档集中的分布情况。本文采用了 TF-IDF 加权方式[15]，这种方式考虑了特征词的局部词频和全局文本词频，计算公式如下：

$$w_{ik}=tf_{ik}.idf_k \tag{1}$$

其中，$tf_{ik}$ 表示特征 $t_k$ 在文本 $d_i$ 中出现的频率；$idf_k$ 表示特征 $t_k$ 的反文档频率，即训练文本集中未出现特征 $t_k$ 的文档比率。常用的计算公式为：

$$idf_k=\log\left(\frac{N}{n_k}+0.01\right) \tag{2}$$

其中，$N$ 表示训练集的总文档数，$n_k$ 表示出现特征 $t_k$ 的文档数量。从该式可以看出，文档集越大，出现特征的文档数量越少，则 $idf_k$ 越大，即该特征对文本类别的区分能力越大。根据此公式得出最终的权重计算公式为：

$$w_{ik}=tf_{ik}.\log\left(\frac{N}{n_k}+0.01\right) \tag{3}$$

考虑到文本长度对权重的影响，需要做归一化处理，将权重进行归一化处理：

$$w_{ik}=\frac{tf_{ik}.\log\left(\frac{N}{n_k}+0.01\right)}{\sqrt{\sum_{i=1}^{n}(tf_{ik})^2.\log^2\left(\frac{N}{n_k}+0.01\right)}} \tag{4}$$

文本 $d_i$ 可以表示成若干词的集合 $d_i=\{t_{i1},t_{i2},\cdots,t_{im}\}$，对每个特征词汇赋予一定的权重$\{w_{i1},w_{i2},\cdots,w_{in}\}$。这样，每个文本都可以映射为一组特征词对应权重组成的向量空间中的一个特征向量[16-17]。对于未在情感词典中出现的文本词汇，可设其分量值为 0。由于单个文本中包含情感词典中的词汇较少，因此用该向量表示的矩阵属于稀疏矩阵。

## 3 基于 PSO-GP 的分类器构建

高斯过程是在贝叶斯框架下对不确定问题进行合理估计的一种统计学习方法，其特征完全由均值函数和协方差函数决定，一般设均值函数为零。因此，协方差函数中超参数的确定成为了影响高斯过程分类性能的重要因素[18]。PSO 算法具有程序化处理方便、参数较少、算法实现简单等特点[8]。本文引入 PSO 进行 GP 超参数优化，以提高高斯过程超参数的寻优速度。

### 3.1 粒子群优化算法

PSO 算法是一种群智能进化算法，它从随机解出发，通过迭代寻找最优解，再通过适应度来评价解的质量，但没有遗传算法的“交叉”(Crossover)和“变异”(Mutation)操作，通过追随当前搜索到的最优值来寻找全局最优。相比传统的遗传算法，PSO 具有算法简单、实现容易、搜索速度快、所含参数较少的特点。不失一般性，本文以最小化系统适应度函数 $f$ 为优化目标，标准粒子群算法可描述为：

$$\begin{cases} v_i^{m+1}=w_i^m v_i^m+c_1r_1(pbest_i^m-x_i^m)+c_2r_2(gbest^m-x_i^m) \\ x_i^{m+1}=x_i^m+v_i^m \\ pbest_i^{m+1}=\begin{cases} x_i^{m+1}, & f(x_i^{m+1})<f(pbest_i^m) \\ pbest_i^m, & f(x_i^{m+1})\geqslant f(pbest_i^m) \end{cases} \\ gbest^{m+1}=\arg\min f(gbest^m),f(pbest_i^{m+1}) \end{cases} \tag{5}$$

其中，$x_j$，$v_i$，$pbest_i$ 分别为第 $i$ 个粒子的当前位置、当前速度、个体历史最优位置；$gbest$ 为所有粒子经历的最好位置；$w$ 为惯性权重，取较大值有利于在局部进行精细搜索；$c_1$ 为个体进化因子，$c_2$ 为社会进化因子，取值区间均为$[0,2]$；$r_1$，$r_2$ 为服从U(0,1)分布的相互独立的随机向量。

### 3.2 基于高斯过程的分类原理

高斯过程最早被用于解决非线性实值预测问题[19]。假定数据集 $D=\{(x_i,y_i),y_i=\pm1,i=1,\cdots,n\}$，其中 $y_i$ 为类标签。二分类问题本质就是在泛函空间 $F$ 中找到一合适的映射 $f(x)$，使得能利用 $y=f(x)$对样本进行正确的分类。基于高斯过程分类的基本思想，就是在以 $f(x)$为高斯过程的假设下，以贝叶斯准则中的后验概率最大化为目标，寻找合适的 $f(x)$。首先假设存在一个隐函数 $f(x)=\phi(x)^{\mathrm{T}}w$ 定义了输入数据和类标签之间的映射关系，同时假定类标签独立同分布，且服从条件概率 $p(y|f(x))$，则对于二分类问题，样本 $x_i$ 属于类别 $y_i$ 的概率为：

$$p(y_i|f_i)=\sigma(y_if_i) \tag{6}$$

其中，$f_i=f(x_i)$为隐函数，$\sigma(\cdot)$为 Sigmoid 类函数，如 Logistic 函数。由于训练样本间相互独立，因此对应的似然概率为：

$$p(y|f)=\prod_{i=1}^{n}p(y_i|f(x_i))=\prod_{i=1}^{n}\sigma(y_if_i) \tag{7}$$

对于样本集 $D=\{(x_i,y_i),y_i=\pm1,i=1\cdots n\}$，高斯过程分类假设 $f(x)\sim GP(0,K)$是一个均值为 0 的高斯过程，则 $p(f|x)$是一个多维高斯密度函数：

$$p(f|x)=\frac{1}{(2\pi)^{\frac{L}{2}}|K|^{\frac{1}{2}}}\exp\{-\frac{1}{2}f^{\mathrm{T}}K^{-1}f\} \tag{8}$$

其中，$K$ 是 $f$ 的协方差矩阵，即核函数。因此，由贝叶斯公式[20]可得后验概率：

$$p(f|y,X)=\frac{p(y|f)p(f|X)}{p(y|X)} \tag{9}$$

其中，$p(y|f)$为似然函数，边缘概率密度为 $p(y|X)=\int p(y|f)p(f|x)\mathrm{d}f$。对高斯过程分类器进行训练，也就是对后验概率$p(f|y,X)$的估计值。因此，给定测试数据点 $x^*$ 的类标签属于正类的概率为：

$$\begin{aligned}\pi^*&=p(y^*=+1|X,y,x^*)\\&=\int\phi(f^*)q(f^*|X,y,x^*)\mathrm{d}f^*\end{aligned} \tag{10}$$

本文可设定当概率 $\pi^*>0.5$ 时为正类样本(+1)，否则将其划分为负类(−1)。

### 3.3 PSO-GP 模型的建立

在分析高斯过程分类机理后，本文充分利用 PSO 在优化超参数方面的优势，构建出 PSO-GP 分类模型，如图 4 所示。

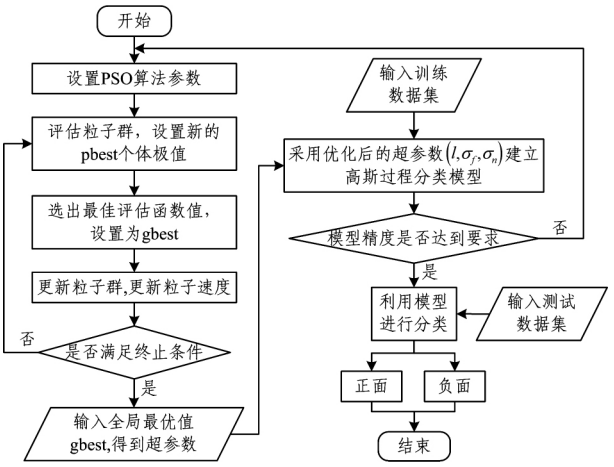



图 4 粒子群算法优化高斯过程模型

具体实现步骤如下：

1)始化 PSO 的网络参数，包括确定粒子群规模、迭代次数、粒子初速度和粒子初始位置。

2)评估粒子群，计算出每个粒子个体的适应度值 $f_i$。

3)将步骤 2)中计算出的适应度值 $f_i$ 与粒子在先前迭代历史中的最佳适应值 $f(pbest_i)$进行比较，若前者小于后者，则用新的适应值取代前轮的 $f(pbest_i)$，用新的粒子取代前轮的粒子。

4)每个粒子的最佳适应度值 $f(pbest_i)$与所有粒子的最佳适应度值 $f(gbest_i)$进行比较。若 $f(pbest_i)<f(gbest_i)$，则用该粒子的最佳适应度值取代原有的全局最佳适应度值，保持粒子当前状态。

5)满足停止条件或者达到迭代次数要求时停止，输出最优值即高斯分类模型的超参数。

## 4 实验结果与分析

### 4.1 实验数据分析

本文将从淘宝(www.taobao.com)、京东(www.360buy.

com)、国美(www.gome.com.cn)和苏宁(www.suning.com)网站上下载的空调评论组成语料库并作为实验数据，语料库中大概包括 150 万个汉字。为了将语料应用于训练和分类，现将语料库均等划分，一半用于高斯分类器的训练，另一半用于分类器的测试语料。为方便标识，用 tb，jd，gm，sn 和 total 分别代表来自淘宝、京东、国美、苏宁以及全部的语料库，并用 1 和 2 区分语料是用于分类还是测试。根据用户打分对空调评论赋标签值，其中 3～5 分正类(+1)，1～2 分为负类(−1)，例如 tb2 即为来自淘宝网的测试语料。测试语料的具体组成如表 1 所列。

表 1　测试语料的组成结构

| 语料 | 总量 | 正类数量(+1) | 负类数量(−1) |
|---|---|---|---|
| tb2 | 4178 | 2993 | 1185 |
| jd2 | 2644 | 1782 | 862 |
| gm2 | 2060 | 1239 | 821 |
| sn2 | 1780 | 1072 | 708 |
| total2 | 10662 | 7086 | 3576 |

### 4.2　分类器的评价指标

本文对分类器的性能进行评测时，采用了微平均(F1)作为评价分类结果的指标，需先计算查准率(precision)和召回率(recall)，方法如式(11)－式(13)[21]所示：

$$precision=\frac{\sum_{c_j\in C} true(c_j)}{\sum_{c_j\in C} doc(c_j)} \tag{11}$$

$$recall=\frac{\sum_{c_j\in C} true(c_j)}{\sum_{c_j\in C} response(c_j)} \tag{12}$$

$$F_1=\frac{2\times precision\times recall}{precision+recall}\times 100\% \tag{13}$$

查准率反映分类器的准确性，查全率反映分类器的完备性。两个评价标准是互补的，单纯提高其中一个评价标准会导致另一个评价标准的降低。微平均 F1 综合了查准率和查全率。在对分类器性能进行评价时，一般都是采用 F1 作为指标。式(11)、式(12)中的 $true(c_j)$ 是分类为 $c_j$ 并且正确的文本数，式(12)中的 $response(c_j)$ 是分类为 $c_j$ 的文本数。为了更加简便地标记各项指标，表 2 列出本文使用的评论指标名称。

表 2　指标表示

| 语料表示评价指标 | P(precision) | R(recall) | F(F1) |
|---|---|---|---|
| P(正面) | 正面查准率 | 正面查全率 | 正面微平均 |
| N(负面) | 负面查准率 | 负面查全率 | 负面微平均 |
| Q(整体) | 整体查准率 | 整体查全率 | 整体微平均 |

### 4.3　实验结果比较

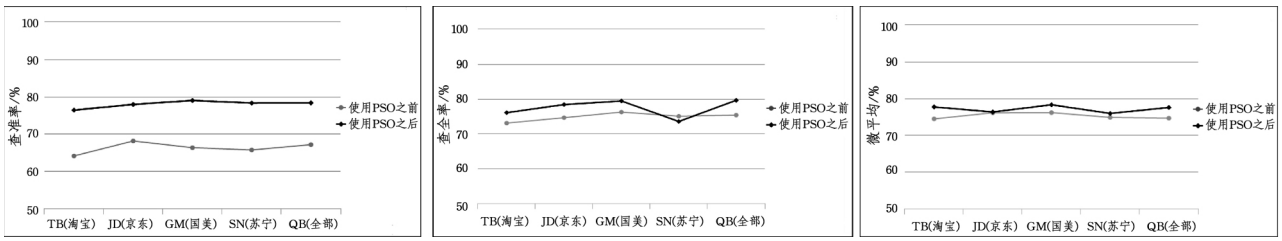
按照前面构建分类器的流程构建分类器，并进行训练，训练完成后分别对测试语料的各部分进行测试，以上过程都在 matlab 下实验。测试结束之后，根据实验结果统计计算查全率、查准率以及微平均等各项指标值。

优化前测试语料各个部分的评价指标如表 3 所列。

表 3　优化前测试语料各个部分的评价指标/%

| 测试语料 | P | | | R | | | F | | |
|---|---|---|---|---|---|---|---|---|---|
| | PP | NP | QP | PR | NR | QR | PF | NF | QF |
| tb2 | 63.21 | 65.74 | 64.08 | 97.65 | 27.91 | 73.04 | 81.32 | 39.07 | 74.42 |
| jd2 | 67.78 | 69.13 | 68.10 | 96.07 | 30.14 | 74.60 | 83.16 | 40.41 | 76.08 |
| gm2 | 65.73 | 67.92 | 66.30 | 97.38 | 31.76 | 76.20 | 82.69 | 42.04 | 76.13 |
| sn2 | 65.40 | 66.31 | 65.68 | 96.09 | 30.81 | 75.01 | 79.37 | 43.67 | 74.82 |
| qb2 | 66.28 | 68.53 | 67.12 | 97.40 | 29.74 | 75.34 | 80.15 | 42.48 | 74.62 |

通过表 3 可以看出，未采用 PSO 优化之前，高斯分类器的准确率仅为 60%左右，微平均指标的值也偏低，因此有必要对高斯过程进行优化。测试使用 PSO 算法优化的高斯过程分类器的分类效果，并与前面的实验数据做对比。在已知粒子的评价函数以及表示方式之后，使用 PSO 算法进行优化。设优化过程中最大的迭代次数为 30，学习因子为 2，粒子种群大小设为 20。每次循环结束后，便可得到一个 gbest 最优极值，根据这个极值计算高斯核函数的超参数，利用这些超参数构建高斯分类模型，并对其进行训练。重复执行上述过程，直至测试数据集的各项评价指标达到最优。最后，根据最优的实验结果统计计算查全率、查准率以及微平均等各项指标值。实验结果统计如图 5(a)－图 5(c)和表 4 所示。




(a)PSO 优化前后查准率的比较　(b)PSO 优化前后查全率的比较　(c)使用 PSO 优化前后微平均的比较

图 5

表 4　优化后测试语料各个部分的评价指标/%

| 测试语料 | P | | | R | | | F | | |
|---|---|---|---|---|---|---|---|---|---|
| | PP | NP | QP | PR | NR | QR | PF | NF | QF |
| tb2 | 76.47 | 76.08 | 76.38 | 98.79 | 30.73 | 76.04 | 85.72 | 42.37 | 77.67 |
| jd2 | 78.37 | 77.13 | 77.90 | 97.30 | 33.28 | 78.36 | 86.46 | 43.61 | 76.30 |
| gm2 | 79.01 | 78.92 | 78.97 | 98.56 | 34.82 | 79.40 | 86.29 | 42.36 | 78.26 |
| sn2 | 78.23 | 78.31 | 78.28 | 97.29 | 33.91 | 73.51 | 84.97 | 44.77 | 75.91 |
| qb2 | 77.94 | 78.53 | 78.32 | 98.27 | 32.64 | 79.60 | 85.05 | 42.08 | 77.52 |

通过图 5 及表 4 可以看出，对高斯过程采用 PSO 算法优化之后，查准率、查全率和微平均 3 项指标均有明显的提高（存在极个别偏低的现象），其中准确率提高了近 15%，已接近 80%，表明采用 PSO 进行参数优化对分类器分类效果的提升有很重要的作用。

结束语 本文实现了基于粒子群优化算法（PSO）优化高斯过程（Guassian Process）的中文文本分类。通过爬虫技术进行语料采集，利用情感词典对中文文本语料进行特征提取，计算特征权重，从而数字化表示特征，最终利用所提分类模型进行中文文本情感分类。实验结果表明，相较于单纯高斯过程分类方法，本文提出的方法较大程度地提高了其性能。

## 参 考 文 献


[1] 王素格，李德玉，魏英杰. 基于赋权粗糙隶属度的文本情感分类方法[J]. 计算机研究与发展，2011，48(5)：855-861.

[2] 马晓玲，金碧漪，范并思. 中文文本情感倾向分析研究[J]. 情报资料工作，2013(1)：52-56.

[3] WANG J X，DONG A. A comparison of two text representations for sentiment analysis[C]// 2010 International Conference on Computer Application and System Modeling (ICCASM 2010). IEEE，2010.

[4] 徐健锋，许园，许元辰，等. 基于语义理解和机器学习的混合的中文文本情感分类算法框架[J]. 计算机科学，2015，42(6)：61-66.

[5] 万源. 基于语义统计分析的网络舆情挖掘技术研究[D]. 武汉：武汉理工大学，2012.

[6] CASALE S，RUSSO A，SCEBBA G，et al. Speech Emotion Classification using Machine Learning Algorithms[C]// The IEEE International Conference on Semantic Computing. Santa Clara，California，USA，August 2008：4-7.

[7] CHANGLI Z，WANLI Z，TAO P，et al. Sentiment Classification for Chinese Reviews Using Machine Learning Methods Based on String Kernel[C]// Third International Conference on Convergence and Hybrid Information Technology. Pusan，Korea，November 2008：11-13.

[8] 王维博. 粒子群优化算法研究及其应用[D]. 成都：西南交通大学，2012.

[9] DAHIWALE P，RAGHUWANSHI M M，MALIK L. Design of Improved Focused Web Crawler by Analyzing Semantic Nature of URL and Anchor Text[C]// 2014 9th International Conference on Industrial and Information Systems. Gwalior，India，December 2014：15-17.

[10] W3C Document Object Model Level 3 Core Specification[OL]. http://www.w3.org/TR/DOM-Level-3-Core.

[11] 湛燕，陈昊，袁方，等. 基于中文文本分类的分词方法研究[J]. 计算机工程与应用，2003，29(23)：87-88，91.

[12] 高丹，张彦霞，赵永恒. 中国虚拟天文台交叉证认工具的开发和应用[J]. 天文学报，2008，49(3)：348-358.

[13] 徐琳宏，林鸿飞，杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报，2007，21(1)：96-100.

[14] 吴艳玲. 基于 SVM 的网页分类器的研究[D]. 长春：吉林大学，2004.

[15] 石慧. 基于特征选择和特征加权算法的文本分类研究[D]. 济南：山东师范大学，2015.

[16] 王之鹏. Web 文本分类系统中文本预处理技术的研究与实现[D]. 南京：南京理工大学，2009.

[17] 平源. 基于支持向量机的聚类及文本分类研究[D]. 北京：北京邮电大学，2012.

[18] PLATANIOS E A，CHATZIS S P. Gaussian Process-Mixture Conditional Heteroscedasticity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2014，36(5)：888-900.

[19] 胡明. 基于高斯过程的非线性预测控制[D]. 广州：华南理工大学，2012.

[20] 王洪春. 贝叶斯公式与贝叶斯统计[J]. 重庆科技学院学报（自然科学版），2010，12(3)：203-205.

[21] 王成，刘亚峰，王新成，等. 分类器的分类性能评价指标[J]. 电子设计工程，2011，19(8)：13-15，21.


（上接第 427 页）


[10] FILIPPONEA M，CAMASTRAB F，MASULLIA F，et al. A survey of kernel and spectral methods for clustering[J]. Pattern Recognition，2008，41(1)：176-190.

[11] LUXBURG U. A tutorial on spectral clustering[J]. Statistics and Computing，2007，17(4)：395-416.

[12] VERMA D，MEILA M. A comparison of spectral clustering algorithms：2003. UW CSE Tech nical report 03-05-01[R]. 2003.

[13] 卜德云，张道强. 自适应谱聚类算法研究[J]. 山东大学学报，2009，39(5)：22-39.

[14] ZHAO F，JIAO L C，LIU H Q，et al. Spectral Clustering with Eigenvector Selection Based on Entropy anking[J]. Neurocomputing，2010，73(10-12)：1704-1717.

[15] 赵凤，焦李成，刘汉强，等. 半监督谱聚类特征向量选择算法[J]. 模式识别与人工智能，2011，24(1)：48-56.

[16] ZELNIK-MANOR L，PERONA P. Self-Tuning Spectral Clustering[C]// Proc of Advances in Neural Information Processing Systems 17. Vancouver，Canada，2004：1601-1608.

[17] NGA Y，JORDAN M I，WEISS Y. On spectral clustering：Analysis and an algorithm[C]// Proceedings of the 14th Advances in Neural Information Processing Systems (NIPS 2002). Cambridge，MIT Press，2002：849-856.

[18] ZELNIK M L，PERONA P. Self-tuning spectral clustering[C]// Advances in Neural Information Processing Systems (NIPS). Cambridge，MA：MIT Press，2004.

[19] 孙继广. 矩阵扰动分析[M]. 北京：科学出版社，2001：146-160.