文章编号:1006 -9348(2013)02 -0299 -04

基于SVM算法的文本分类技术研究

崔建明[1],刘建明[2],廖周宇[2]
(1. 桂林理工大学现代教育与技术中心,广西 桂林 541004;
2. 桂林理工大学信息科学与工程学院,广西 桂林 541004)

摘要: 在优化分类技术的研究中,文本特征化后通常具有高维性和不平衡性的特点,导致传统的分类算法准确率不高的问题。针对文本分类器的性能容易受到核函数和参数的影响的问题,为提高文本分类器的准确性。采用支持向量机(SVM)的理论在文本分类技术同时将根据优化的粒子群算法(PSO)引入SVM分类算法中进行优化文本分类器的参数,将分类器的准确率作为PSO算法适应度函数通过粒子移动操作找出最佳参数并用SVM算法进行分类。在文本数据集上的仿真结果表明,与传统的算法相比,经PSO算法优化后的SVM文本分类器的准确性更高,PSO算法是一种有效的优化方法,能广泛应用于文本分类问题。

关键词: 支持向量机; 文本分类; 算法
中图分类号: TP391.9 **文献标识码:** A

Research of Text Categorization Based on Support Vector Machine

CUI Jian - ming[1], LIU Jian - ming[2], LIAO Zhou - yu[2]
(1. Center of Modern Education and Technology, Guilin University of Technology;
2. School or College of Information science and engineering, Guilin University of Technology, Guilin Guangxi 541004, China)

ABSTRACT: Text characterization usually has the characteristics of high dimensional and unbalanced, which causes the probems that traditional classification algorithm accuracy is not high, the performance of text categorization is vulnerable to the influence of kernel function and parameters. In order to improve the accuracy of the text classifier, this article used the support vector machine (SVM) theory to study the text classification technology, and the theory of particle swarm optimization (PSO) algorithm, the classification algorithm was introduced to the SVM to optimize the parameters of the text classifie, we used the accuracy of the classifier as fitness functions, used particles move operation to find the best parameters, and used the SVM algorithm to classify the texts. Compared with the traditional algorithm, the new classifier has higher accuracy.

KEYWORDS: SVM; Text categorization; Algorithm

1 引言

网络技术的不断发展,互联网成了人们获取信息的重要途径,但因特网上的信息以爆炸式的增长,而且网络信息是没有次序的,因而人们很难准确而有效的获取需要的信息。面对如此庞大而且不断增长的信息,如何有效地组织并找到用户需要的信息是当代信息科学技术领域的一大难题,因此,如何使得信息文本分类是机器学习中一个课题,应用机器学习实现按照文本内容自动分类技术是解决信息准确、快速检索的主要方法之一[1,2]。

目前,应用于文本分类的技术和算法很多,例如有朴素贝叶斯算法、K最近邻算法、神经网络、支持向量机(Support Vector Machine 即SVM)等[3]。其中,SVM分类算法有很好的泛化能力与学习能力,SVM分类算法是以结构风险最小化为目标,所求得的解是全局最优解,该算法克服“维数灾难”问题。有比较深厚的理论基础,被广泛应用于文本自动分类、人脸识别、基因表达、手写体的识别等领域。

对于传统的SVM分类算法,易受数据集,分类器及训练参数的影响,本文针对训练参数对SVM分类器准确率影响,可以利用基于优化理论的粒子群算法对核函数参数和分类器参数进行优化。粒子群优化算法是一种基于群体的全局

收稿日期:2012 -04 -20 修回日期:2012 -05 -27

优化算法，该算法具有自我学习和向他人学习的优点，能在较少的迭代中找到全局最优解。利用 PSO 算法对支持向量机分类器进行优化后，最后利用文本数据进行试验仿真，与传统的 SVM 算法相比，验证算法的有效性。

2 文本分类技术

2.1 文本分类描述

早在互联网流行前，人们就已经开始了对文档自动分类算法的研究。文本自动分类为[4]：在确定的分类目标下，将待分类的网页文本根据内容自动的划分到某个类别中，使得网页文本具有正确的标签。文本分类是个有监督学习的过程，以已标注的文本集为基础，通过分类器找出文本类别与文本特征之间的关系，然后利用这个关系模型对新的文本进行类别的判断，文本分类其实是将未标明的文本映射到预先确定文本版类别的集合，该映射可以是一对一或是一对多。用数学公式表示如下：

$$f: A \rightarrow B$$

其中，$A=\{D_1, D_2, \cdots, D_n\}$，$B=\{C_1, C_2, \cdots, C_m\}$。$A$ 是需要分类的文本样本数据集，B 是文本样本所属分类类别的集合。文本分类就是根据分类的准建立相应的判别函数公式；遇到新的文档时，根据所得判别公式，确定文本所属的类别。

2.2 文本分类过程

本质上，文本分类是一个文本模式特征进行识别过程，文本分类由训练和测试两部分组成[5]。两部分是独立的过程。训练过程是为分类器提供测试的一个过程，而测试结果也会反作用于分类器，同时调整结果形成新的分类器；分类过程即是使用学习过程所得到得分类器对待分类的文本进行分类，输出待测文本所属类别。文本分类系统原理如图 1 所示。

图 1 文本分类系统原理图

2.3 文本表示技术

VSM（向量空间模型）最早是 Gerard Salton 等人在 1958 年提出的，最早应用于信息检索领域，著名的 SMART（System for the Manipulation and Retrieval of Text）系统就应用了向量空间模型的技术，后来在文本分类得到了广泛的应用，该模型用 d（document）表示文本，用 t（term）表示特征项，用 w（weight）表示权重。因此，在向量空间模型中文本被表示为 $D=(t_1, w_1; t_2, w_2; \cdots, t_n, w_n)$，其中 $t_1, t_2, \cdots, t_n$ 与特征词的词典对应，t 在文本中有先后的顺序，但目前的文本自动分类都不考虑顺序，把 $t_1, t_2, \cdots, t_n$ 看成 n 维空间的坐标系，$w_1, w_2, \cdots, w_n$ 看成对应 n 维空间的坐标值，则 $D=(t_1, w_1; t_2, w_2; \cdots, t_n, w_n)$ 就是 n 维空间向量。

两个文本之间的相关程度可以用它们之间的相似度来度量，相似度计算有多种计算公式常用的有欧氏距离公式，向量内积公式，夹角余弦公式等。

3 支持向量机理论

SVM 算法是 Vapnik 和其领导的贝尔实验室小组在 1995 年提出[6,7]的一种基于统计学习理论的新型的通用学习方法，它是基于统计学习理论的 VC 理论和结构风险最小化原理的基础上发展起来的。SVM 的基本原理为：假设存在训练样本$\{(x_i, y_i)\}$，$i=1,2,\cdots,m$，可以被某个超平面 $w \cdot x + b = 0$ 没有错地分开，其中，$x_i \in R^n$，$y_i \in \{-1, 1\}$，m 为样本个数，R^n 为 n 维实数空间。因此和两类最近的样本点距离最大的分类超平面称为最优超平面，如图 2 所示，H 为最优超平面。最优超平面只和离它最近的少量样本点即支持向量确定。

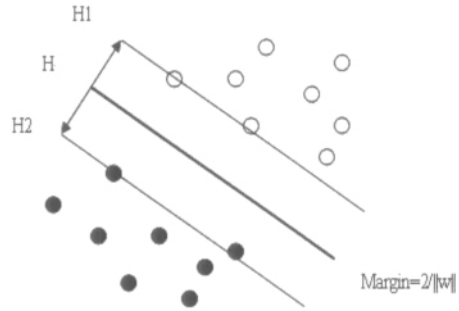

图 2 SVM 原理

图 2 中的空心圆点和黑心圆点代表两类不同类别的样本；H 为分类线，$H1$，$H2$ 分别为平行于分类线的直线，它们经过离分类线最近的那些少量的样本点，两者间距离称为分类间隔。H 线将两个不同的类正确隔离，同时使分类间隔最大化。设样本集为(x_i, y_i)，$i=1,2,\cdots,m$，$y_i \in \{-1, 1\}$，并满足：

$$y_i [(w \cdot x_i) - b] - 1 \geqslant 0 \qquad (1)$$

该分类的间隔等于$\frac{2}{\|w\|}$，其间隔最大等价于求$\|w\|^2$的值最小。满足上式（1）且$\frac{1}{2}\|w\|^2$最小的分类平面叫做最优分类面，$H1$，$H2$ 两条平行直线上的那些训练样本点称为支持向量。

利用拉格朗日方法可以把求解最优分类面的原始问题转化为其对偶问题，即在条件 $\sum_{i=1}^{n} a_i y_i = 0$，$a_i \geqslant 0$，$i = 1,2,3,\ldots n$ 下对 a_i 求解以下最大的函数值：

$$Q(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j (x_i, x_j) \qquad (2)$$

a_i 为原问题中与（1）式对应的拉格朗日乘子。该问题就是求解二次凸规划的优化问题，存在唯一最优解，可以证明，有些

乘子 a_i 不为零，它也就是支持向量。求解该问题得到最优平面的 w^* 和 b^*，此时最优分类函数为

$$D(x) = \mathrm{sgn}((w^* \cdot x) - b^*) = \mathrm{sgn}\left(\sum_{i,j=1}^{n} a_i^* y_i (x_i \cdot x) - b^*\right) \quad (3)$$

求和只对支持向量进行；b^* 是偏移量。根据泛函分析中的度量空间理论，倘若有一种核函数 $k(x_i, x_j)$ 满足 Mercer 条件的核函数替换线性算法的内积就可以找到原输入空间中对应的非线性算法。如果用特征空间的 $\varphi(x)$ 代替 x，则式(3)转化为

$$Q(a) = \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j \varphi(x_i)\varphi(x_j) \quad (4)$$

而相应的分类函数变为

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} a_i^* y_i k(x_i \cdot x) - b^*\right) \quad (5)$$

在 SVM 分类算法中，如果定义不同的内积函数，就能实现多项式逼近、贝叶斯分类器、径向基函数(RBF)方法等选用不同的核函数就可以构造不同的 SVM[3]。

4 基于 SVM 文本分类算法

4.1 PSO－SVM 文本分类器设计

基于统计学习方法 SVM 文本分类算法，其最终的目的是找出能把文本训练样本进行无误划分的最优超平面。对文本进行分类时，首先是从具体问题中获取文本训练数据进行预处理即利用向量空间模型把文本表示成向量形式，其次就是根据具体的需要选择恰当的核函数及核函数参数，该过程利用 PSO 算法求解，其优化方法模型如图 3 最后利用求得最优参数进行训练文本样本。

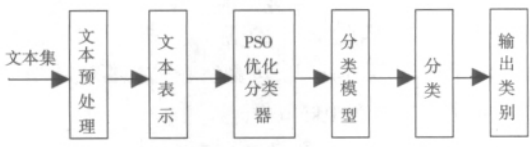

图 3 PSO 优化方法模型

SVM 文本分类算法步骤如下:

1) 利用向量空间模型处理方法把文本数据转化为 SVM 分类算法能处理的形式;

2) 选择合适核函数，众多实验表明，一般情况下选择 RBF 作为核函数所得结果最好。

3) 求解最优的参数。利用 PSO 最优化算法找出 SVM 分类器的最优参数。

4) 利用 3) 所得到的最优参数应用 SVM 算法分类器来对文本样本数据进行训练并用测试集进行分类预测实验。

4.2 SVM 分类算法中所选核函数参数优化

粒子群优化算法(PSO)是群体智能优化的算法，起源于人们对鸟类捕食行为的学习，该算法利用生物种群行为特征求解最优化问题。算法中一个粒子代表一个问题的潜在解，每个粒子对应一个有适应度函数决定的适应度值。利用 PSO 对 SVM 文本分类算法寻优的步骤如下[8,9]:

1) 输入经过预处理的含有特征的训练文本样本;

2) PSO 算法以及 SVM 的核函数参数初始化;

3) 利用随机函数方法初始化种群的速度和粒子，以 SVM 算法所求得的准确率率作为粒子的适应度;

4) 对 PSO 算法中的种群个体通过粒子更新，产生新的粒子，并计算新种群粒子的适应度值。

5) 判断当前的粒子的个体极值是否为种群的全局最优解，是，就找到全局最优子集，若否，继续上一步的循环操作;

6) 将优化后的核函数参数利用 SVM 文本分类器训练，并用文本测试集进行测试。PSO 优化算法核函数参数优化过程图如图 4 所示。

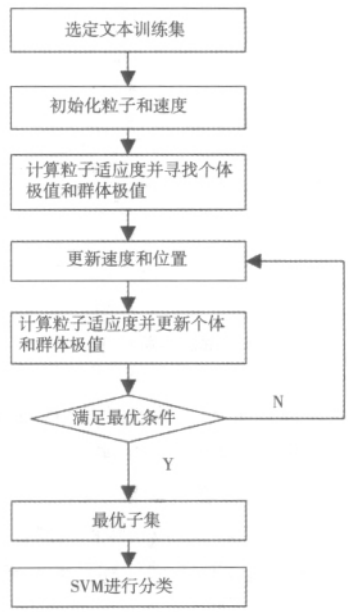

图 4 PSO 算法优化 SVM 分类器流程图

5 仿真研究

5.1 实验环境与数据来源

实验数据来源于搜狗分类新闻语料库和 20 组新闻数据(经典的文本分类数据集)。数据预处理的特征词选择方法为 IG(信息增益)。实验数据包含 150 个文本特征属性。样本数据为 2000，其中 1000 为训练集，1000 为测试集，数据分新闻，非新闻类两类。本实验平台基于 MATLAB 7.6 与开源软件 LIBSVM，LIBSVM 工具箱是台湾大学林智仁教授等研发的非常简单而且有效的 SVM 模式识别与回归软件包。本实验利用 LIBSVM 下的 MATLAB 接口函数。

5.2 结果分析

实验中影响分类器性能的因素很多，但在测试过程中，

本文以准确率作为评估文本分类器性能的方法。准确率 P 的计算方法为:

$$P = (\text{正确划分样本数} / \text{样本总数}) * 100\%$$

为了能直观的观察测试样本数据,并能查看 SVM 算法中的 SV 样本,在 MATLAB 中画出各个数据样本点在二维空间图中分布情形,其训练文本样本点的二维分布效果图如图 5 所示。图中的 + 号为新闻类,* 号为非新闻类,圆圈表示的支持向量。观察发现样本的分布不是很平衡,属于新闻类的样本数占大多数,这与现实中样本的分布不平衡也相符合的。文本的特征词选择算法是信息增益方法,信息增益基于信息论原理通过统计某个词项在一篇文档中出现与否来预测文本的类别,是一种广泛的特征选择方法。

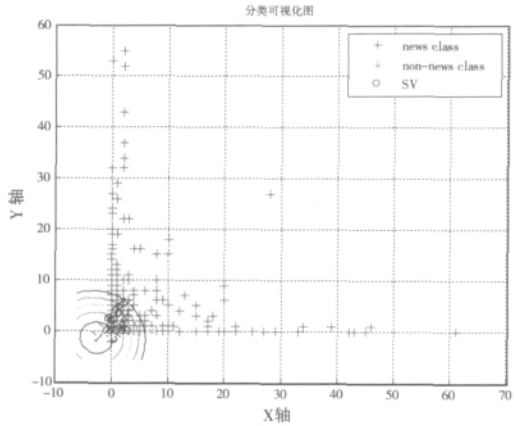

图 5　样本可视化图

利用 PSO 算法优化文本分类器时,初始的种群为 20,进化代数为 150。PSO 算法寻优的准确率变化曲线图如图 6 所示。图 5 中选取了 PSO 算法迭代数中第 1、30、60、90、120、150 代的最佳准确率和平均准确率进行了分析。图 5 中的两条变化曲线,分别为最佳准确率曲线和平均准确率曲线。

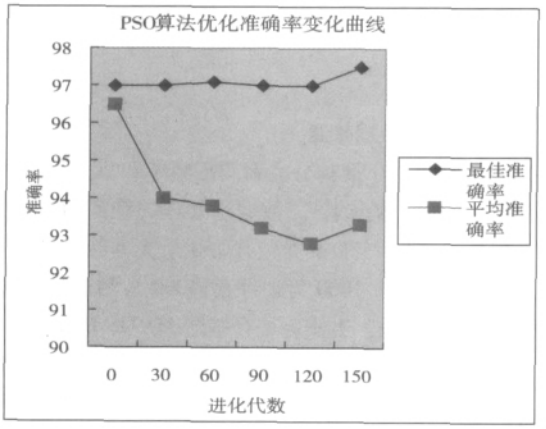

图 6　迭代进化过程图

为验证本文算法的有效性,在实验过程利用同一数据集,运用传统的 SVM 算法进行训练和测试,同时与本文的方法(PSO 算法优化的 SVM 算法)的实验结果进行对比分析。传统 SVM 的采用默认参数: c = 1,g = 1/k,其中 k 为特征属性数,k = 150。基于 PSO 算法优化的得到参数: c = 0.1,g = 0.0839。PSO 算法优化 SVM 文本分类的数据分析结果如表 1 所示。

表 1　PSO 算法优化前后的准确率对比表

数据集	参数 | P	SVM	PSOSVM
sogounews	c,	c = 1	c = 0.1
	g	g = 0.0067	g = 0.0839
	准确率 P(%)	95.4327	97.5000
news20	c,	c = 1	c = 0.1
	g	g = 0.0067	g = 0.0839
	准确率 P(%)	74.022	82.2032

由表 1 可知,在不同数据集上测试时,优化后的 SVM 均比传统的 SVM 算法表现更好其分类器的准确率更高。本文采用的文本特征选择算法为信息增益法,信息增益选择方法会存在数据稀疏问题,因此,对于样本分布不均匀的时,分类的效果有所差别即不同的数据集的分类效果有所不同。通过实验对比可知,应用 PSO 算法对文本的分类器进行优化后比传统的 SVM 分类的效果更好,该方法具备一定的实用性。

6　结束语

通过上述对比,本文从 SVM 的原理出发讨论其在文本分类中的应用,文章讨论了 SVM 分类算法在参数选择上的不足,因此,本文利用 PSO 算法进行了参数优化,通过仿真对比实验,实验数据表明,利用 PSO 算法优化后,文本分类的准确率有一定的提高,能从局部上改善 SVM 分类算法性能。但文章只是在 SVM 算法的基础上进行优化,并未对 SVM 算法理论进行改进,因此,本文的下一工作为将一种新型 SVM(孪生 SVM)引入文本分类,该算法在处理像文本分类这样的不均衡数据问题有很好效果。

参考文献:

[1] 冯是聪,张志刚,李晓明. 一种中文网页自动分类方法的实现及应用[J]. 计算机工程,2004,30(5):19-21.

[2] 汪光庆. 基于 SVM 的网页分类技术研究[D]. 中国石油大学硕士学位论文,2011.

[3] 牛强,王志晓,陈岱. 基于 SVM 的中文网页分类方法的研究[J]. 计算机工程与应用,2007,28(8):1893-1895.

[4] B Scholopf, et al. Estimating the support of a high-dimensional distribution[J]. Neural Computation, 2001,13(7):1443-1472.

[5] 林士敏,田凤占,陆玉吕. 贝叶斯学习、贝叶斯网络与数据采掘[J]. 计算机科学,2005,27(10):69-72.

[6] 边肇棋,张学工. 模式识别(第二版)[M]. 北京:清华大学出版社,2002.

[7] C J Burges. A tutorial on support vector machines for Pattern recognition[J]. Data Mining

(下转第 368 页)

4 实验结果

已经在配置为 Intel(R) XEON(R) CPU，NVIDIA GeForce GTX 470，Windows 7 OS 的电脑上实现了上述算法。为简便起见实验中渲染的对象为长方体，其包含的粒子数为 16，统一网格的大小为 100^3。实验中，为刚体分配的纹理大小为 128^2，为粒子分配的纹理大小为 512^2，为统一网格分配的纹理大小为 1024^2。实验结果如图 8。

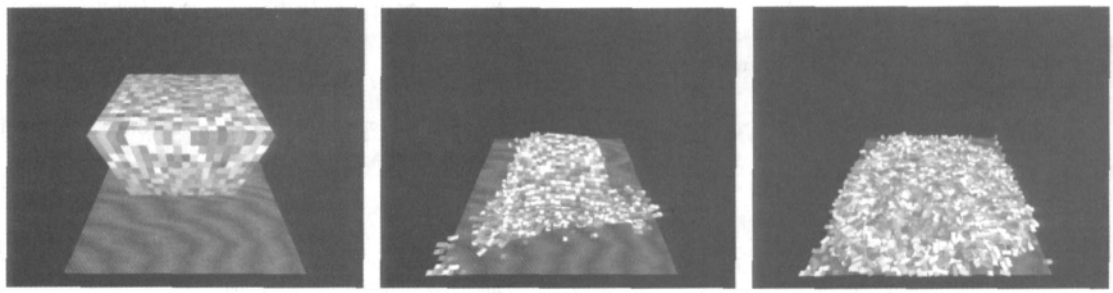

图 8　本实验中包含长方体 6000 个，粒子 96000 个，仿真速率为 238 帧每秒

5 总结

本文使用统一网格的方式来离散化仿真域，当仿真域比较大时，用来保存粒子索引的 2D 纹理也会比较大，而且刚体不会均匀的分布于整个仿真域，使得很多的体素没有包含粒子，导致空间的浪费。在未来的工作中，将会对该问题展开研究，希望能使用一种动态的数据结构来离散化仿真域。

参考文献:

[1] Nathan Bell，Yu Yizhou. Particle - based simulation of granular materials [J]. ACM SIGGRAPH/Eurographics Symposium on Computer Animation. 2005，198 - 8.

[2] Rahul Narain，Abhinav Golas. Free Flowing Granular Materials with Two - Way Solid Coupling [J]. ACM SIGGRAPH Asia. 2010，29 - 6.

[3] Cundall P A Strack. A discrete numerical model for granular assembilies [M]. Geotechnique，1979，29 - 1.

[4] C S Campbell. Rapid granular flows [J]. Annu. Rev. Fluid Mech. 1990，22: 57 - 92.

[5] B K Mishra，C V R Murty. On the determination of contact parameters for realistic DEM simulations of ball mills [J]. Powder Technology，2001，115 - 3.

[6] Frank Losasso. Simulating Water and Smoke with an Octree Data Structure [J]. ACM SIGGRAPH，2004，23 - 3.

[7] B K Mishra. A review of computer simulation of tumbling mills by the discrete element method: Part I—contact mechanics [J]. Int. J. Miner. Process，2003，71 - 1.

[8] J M Ting，B T Corkum. A computational laboratory for discrete element geomechanics [J]. J. comput. Civ. Eng. ASCE. 1992，6 - 2.

[9] 周衍柏. 理论力学教程(第二版) [M]. 北京: 高等教育出版社，1986: 155 - 183.

[10] 仇德元. GPGPU 编程技术: 从 GLSL，CUDA 到 OpenCL [M]. 北京: 机械工业出版社，2011: 67 - 110.

[作者简介]

彭林春(1987 -)，男(汉族)，重庆市人，硕士研究生，主要研究领域为计算机图形学。

杨红雨(1967 -)，女(汉族)，四川成都市人，教授，博士生导师，主要研究领域是计算机图像处理和图形学，计算机仿真和实时软件工程。

杨　光(1974 -)，男(汉族)，陕西西安人，工程师，主要研究领域为流量管理，空域管理，模拟机深度开发。

(上接第 302 页)

and Knowledge Discovery，1998，2(2) : 121 - 167.

[8] 冯是聪，张志刚，李晓明. GA_SVM 在文本分类算法中应用研究 [J]. 计算机仿真，2011，28(1) : 222 - 225.

[9] 史峰，王小川，郁磊，李洋. MATLAB 神经网络 30 个案例分析 [M]. 北京: 北京航空航天大学出版社，2010.

[作者简介]

崔建明(1962 -)，男(汉族)，广西玉林人，副教授，研究方向为网络与数据库技术、云计算。

刘建明(1986 -)，男(汉族)，广西玉林人，硕士研究生，研究方向为数据挖掘技术。

廖周宇(1985 -)，男(汉族)，四川成都人，硕士研究生，研究方向为云计算。