DOI:10.16652/j.issn.1004-373x.2018.15.041

基于支持向量机的不均衡文本分类方法

高　超，许翰林
（南京信息工程大学 电子与信息工程学院，江苏 南京　210044）

摘　要：目前支持向量机(SVM)对均衡文本数据集进行文本分类时表现十分良好,但如果文本数据集是不均衡的,尤其是当不均衡率很大时,容易导致支持向量机分类失败。提出PSO-SMOTE混合算法,针对不均衡文本数据集问题,运用SMOTE算法生成插值样本均衡数据集,并通过PSO算法迭代进化得到最佳的插值样本,对支持向量机的文本分类能力进行优化。实验结果表明,新算法大幅优化了支持向量机分类不均衡文本数据集的能力。

关键词：混合算法；支持向量机；不均衡数据集；插值样本；文本分类；迭代进化

中图分类号：TN911.1-34；TP391.9　　**文献标识码**：A　　**文章编号**：1004-373X(2018)15-0183-04

Unbalanced text classification method based on support vector machine

GAO Chao，XU Hanlin
(School of Electronic & Information Engineering，Nanjing University of Information Science and Technology，Nanjing 210044，China)

Abstract：The support vector machine (SVM) performs well in text categorization of balanced text datasets，but will cause the classification failure when the text dataset is unbalanced，especially for high unbalanced ratio. The PSO-SMOTE hybrid algorithm is proposed to solve the unbalanced text datasets. The SMOTE (synthetic minority oversampling technique) algorithm is used to generate the balanced dataset of interpolation sample，and then the iterative evolution is performed for the interpolation sample by means of PSO algorithm to obtain the optimal interpolation sample，and optimize the text classification performance of SVM. The experimental results show that the new algorithm can greatly optimize the ability of SVM to classify the unbalanced text datasets.

Keywords：hybrid algorithm; support vector machine; unbalanced dataset; interpolation sample; text classification; iterative evolution

0　引　言

通常来说,文本分类的主要任务是将未被标记的文档自动分类到预定义的类别。常见的文本分类方法有支持向量机(Support Vector Machine,SVM)、K最近邻算法(KNN)和朴素贝叶斯(Native Bayes)等。学术界的许多学者也针对这些算法做出了改进研究。文献[1]提出一种基于聚类的改进KNN算法。文献[2]提出一种加权补集贝叶斯文本分类算法。

与其他的文本分类方法相比,支持向量机因为其强大的理论背景以及在处理分类问题中所表现出的优异的泛化能力,越来越多的学者倾向于使用支持向量机进行文本分类。虽然支持向量机十分擅长对均衡文本数据集进行文本分类,但在面对不均衡文本数据集时的分类表现却差强人意。本文针对这一问题提出SMOTE-PSO混合算法对支持向量机进行优化。混合算法生成的最优插值样本均衡了数据集,改善了支持向量机分类不均衡文本集的表现。

收稿日期:2017-11-14　　**修回日期**:2017-12-18

1　支持向量机介绍

支持向量机方法建立在统计学习理论的VC维理论和结构风险最小原理基础上,根据有限的样本信息在模型的复杂性(即对特定训练样本的学习精度,Accuracy)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷,以期获得最好的推广能力(或称泛化能力)。支持向量机的基本原理为: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$\omega\cdot x+b=0$无错地分开,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。因此与两类最近的样本点距离最大的分类超平面为最优超平面。如图1所示,H为最优超平面。最优超平面只由离它最近的少量样本点即支持向量确定。

支持向量机转化成数学形式为一个带约束的最小值问题:

$$\min \frac{1}{2}\|\omega\|^2 \quad \text{s.t.} \quad y_i(\omega \cdot x_i + b) \geqslant 1, \quad i = 1,2,\cdots,n \tag{1}$$

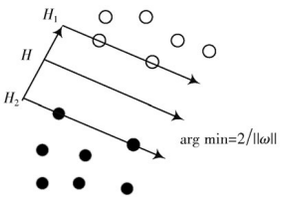

图1 支持向量机原理图

Fig. 1 Schematic diagram of support vector machine

2 PSO算法

粒子群优化算法(PSO)是一种基于种群信息共享概念的最优解搜索方法。粒子群算法首先对种群中的粒子进行初始化,随机分配它们的位置。种群中的每一个个体(粒子)在搜索空间中根据之前的搜索经验不断改进自己的搜索位置。

在搜索最优解的过程中,粒子不断调整它们的位置和速度。将每一个粒子搜索到的历史最优值设为$P_b(t)$,全部粒子搜索到的最优值被称作全局历史最优解,设为$P_g(t)$。

粒子P_i的速度更新公式如下:

$$V_i(t+1) = wV_i(t) + c_1 \cdot r_1(t)\left(P_b(t) - P_i(t)\right) + c_2 \cdot r_2(t)\left(P_g(t) - P_i(t)\right) \tag{2}$$

式中:w为惯性权重;c_1是粒子跟踪自己历史最优值的权重系数;c_2是粒子跟踪群体最优值的权重系数;r_1和r_2是区间[0,1]的随机数。

粒子P_i的位置更新公式如下:

$$P_i(t+1) = P_i(t) + V_i(t) \tag{3}$$

3 SMOTE算法

SMOTE(Synthetic Minority Oversampling Technique)算法的基本思想是在少数类中相距较近的样本之间插入一个人工合成的样本均衡数据集。算法的具体步骤如下:对少数类中的每一个样本x_i,寻找k个与x_i距离最近的样本点。在k个样本点中随机抽取n个样本点$x_{ij}(j = 1,2,\cdots,n)$与x_i进行线性插值操作,生成插值样本p_j。

$$p_j = x_i + \text{rand}(0,1) \times (x_i - x_{ij}) \tag{4}$$

式中rand(0,1)表示区间[0,1]中的一个随机数。插值样本生成示意图如图2所示。

4 基于支持向量机的文本分类算法优化

4.1 PSO-SMOTE文本分类器设计

本文提出的PSO-SMOTE文本分类器与传统的支持向量机文本分类器相比,对不均衡文本集的分类效果更佳。首先对文本数据集进行预处理,将数据集转换为支持向量机可以处理的形式。需要利用向量空间模型(Vector Space Model)把文本表示成向量形式。在向量空间模型中,文本$d_k = (w_{1,k}, w_{2,k}, \cdots, w_{i,k}, \cdots, w_{n,k})$。其中,$w_{i,k}$代表文本$d_k$中特征项$t_i$的权重。文本被表示成向量形式之后往往会因为向量维数过高影响分类效果。本文选择CHI(卡方统计量)作为文本特征选择方法进行降维处理。使用本文提出的PSO-SMOTE算法在少数类中生成最优插值样本均衡文本数据集。最终使用支持向量机可以找出有效的超平面,利用分类模型对文本数据集进行有效地划分。PSO-SMOTE优化分类器模型如图3所示。

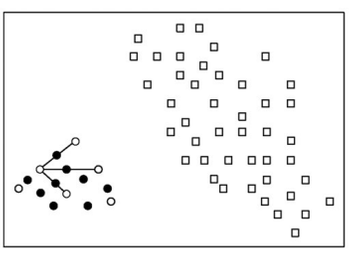

图2 插值样本生成

Fig. 2 Generation of interpolation sample

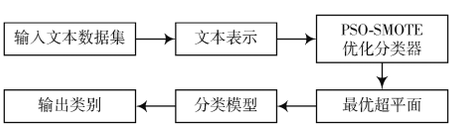

图3 PSO-SMOTE优化分类器模型

Fig. 3 Model of PSO-SMOTE optimizing classifier

4.2 PSO-SMOTE算法生成最优插值样本

通过原始的SMOTE算法简单的生成随机插值样本很有可能产生噪音插值样本对分类结果造成影响[3],所以需要对新生成的插值样本进行一定的筛选。本文提出的PSO-SMOTE混合算法可以简单有效地得到最优的插值样本,提高支持向量机文本分类的表现。PSO-SMOTE算法步骤如下:

1)输入经过预处理的训练文本数据集,获取少数类样本数据集合。

2)根据式(6)生成随机插值样本,初始化大小为$m \times q \times r$粒子群。其中,r是每个插值样本的维度,q是插值样本的个数,m是种群中粒子的个数。

3)初始化粒子群的速度$V_i, i = 1,2,\cdots,q \times r$。

4)对由插值样本组成的粒子$P_i, i = 1,2,\cdots,m$中的每一个位置P_i用支持向量机分类算法进行训练并用适应度函数计算它的适应度值。

5)初始化每一个粒子的最优位置为它的初始位置,$P_{bi} = P_i, i = 1,2,\cdots,m$。

6）得到种群中的全局最优粒子P_g。分别根据式(2),式(3)更新每一个粒子的速度、位置。

7）用支持向量机分类算法训练每一个候选粒子并计算适应度值。

8）如果在这个粒子当前位置计算出更小的适应度值,则更新这个粒子的历史最优值P_b。

9）如果设置的进化停止条件还没满足,则返回步骤6)继续循环;如果已经满足停止条件,则得到全局历史最优粒子。组成这个粒子的插值样本即为全局最优插值样本。

4.3　PSO-SMOTE适应度函数的选择

适应度函数提供了找出最优解的方式并且掌控着粒子的进化过程。适应度函数帮助PSO算法评估每一个候选粒子即每一个问题的潜在解的优劣,所以选择一个合适的适应度函数是非常重要的。本文选取G-mean作为适应度函数。G-mean的计算公式如下:

$$\text{G-mean}=\sqrt{\frac{T_p}{T_p+F_N}\cdot\frac{T_N}{T_N+F_p}} \tag{5}$$

式中:T_p代表正类样本最终被预测分类为正类的样本个数;T_N代表负类样本最终被预测分类为负类的样本个数;F_p代表负类样本最终被预测分类为正类的样本个数;F_N代表正类样本最终被预测分类为负类的样本个数。学术界学者通常会利用G-mean来度量分类器分类不均衡数据集的能力。

5　实验与结果分析

5.1　实验数据

本文从搜狗实验室中选取编辑经过手工整理的新闻语料,并且已经分类为经济、社会、体育、环境、政治五大类。本实验的重点是测评PSO-SMOTE混合算法优化支持向量机分类不均衡文本的能力,所以刻意将每类新闻文本与其他非新闻文本数据混合构成不均衡文本数据集。文本数据的文本特征选择采用卡方统计量算法。如表1所示,为了提升实验数据的复杂度,每一类不均衡文本集的不均衡率都不相同。

表1　不均衡文本数据集

Table 1　Dataset of unbalanced text

文本数据集	新闻类文本	非新闻类文本	文本特征数	不均衡率
经济	156	584	150	1∶0.374 3
社会	715	2 489	150	1∶0.348 1
体育	145	300	150	1∶0.207 0
环境	77	436	150	1∶0.566 2
政治	269	628	150	1∶0.233 4

5.2　实验结果

如图4所示,实验选取了一个不均衡文本集并采用本文所提出的PSO-SMOTE算法对在少数类中生成的插值样本进行迭代优化。初始化种群中的粒子数为30,进化迭代次数设为200,适应度函数选取G-mean。群体中的每一个粒子由随机生成的插值样本组成。图4中的两条曲线分别为群体的最佳G-mean曲线以及平均G-mean曲线。可以明显看出,随着进化迭代次数的增加,种群中的粒子也在不断优化。证明本文提出的PSO-SMOTE算法可以有效地对经典SMOTE算法在不均衡文本集中生成的随机插值样本进行优化。

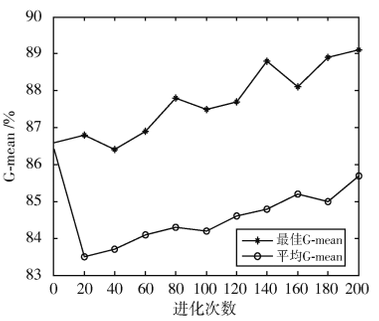

图4　插值样本进化过程图

Fig. 4　Evolution process of interpolation sample

为了突出本文提出的PSO-SMOTE算法对支持向量机分类不均衡文本数据集的优化,实验将PSO-SMOTE算法与SMOTE算法以及经典支持向量机算法作为对比。分别对5.1节中整理的不均衡文本数据集进行分类运算,性能评价标准依旧选择G-mean。实验过程中选择支持向量机的核参数为RBF核,采取交叉验证的方式确定惩罚系数C和核宽度σ的值。实验结果如表2所示。

从表2可以看出,PSO-SMOTE算法改进了支持向量机分类不均衡文本数据集的能力。证明此算法存在一定的实用性和推广价值。

表2　不同算法对不均衡文本集分类能力对比表

Table 2　Comparison of classification ability of unbalanced text set with different algorithrns

文本数据集	SVM	SMOTE	PSO-SMOTE
经济	0.724	0.781	0.863
社会	0.751	0.825	0.884
体育	0.813	0.891	0.965
环境	0.729	0.778	0.856
政治	0.814	0.848	0.889

6　结　语

本文针对支持向量机在文本分类中分类不均衡文本数据集的局限,提出PSO-SMOTE混合算法优化支持向量机的文本分类能力。实验结果表明,本文提出的混合算法有效地提升了支持向量机分类不均衡文本数据

集的能力。下一步的工作方向是对支持向量机的理论进行优化,将改进的支持向量机与PSO-SMOTE算法相结合进一步提升分类能力,并对PSO-SMOTE算法进行进一步改进,尝试利用支持向量生成插值样本。

参 考 文 献

[1] 周庆平,谭长庚,王宏君,等.基于聚类改进的KNN文本分类算法[J].计算机应用研究,2016,33(11):3374-3377.
ZHOU Qingping, TAN Changgeng, WANG Hongjun, et al. Improved KNN text classification algorithm based on clustering [J]. Application research of computers, 2016, 33(11): 3374-3377.

[2] 杜选.基于加权补集的朴素贝叶斯文本分类算法研究[J].计算机应用与软件,2014,31(9):253-255.
DU Xuan. Research on weighted complement-based naive Bayes text classification algorithm [J]. Computer applications and software, 2014, 31(9): 253-255.

[3] 陈斌.SMOTE不平衡数据过采样算法的改进与应用[D].南宁:广西大学,2015.
CHEN Bin. The improvement and application of SMOTE algorithm for unbalanced data sampling [D]. Nanning: Guangxi University, 2015.

[4] 崔建明,刘建明,廖周宇.基于SVM算法的文本分类技术研究[J].计算机仿真,2013,30(2):299-302.
CUI Jianming, LIU Jianming, LIAO Zhouyu. Research of text categorization based on support vector machine [J]. Computer simulation, 2013, 30(2): 299-302.

[5] 谢娜娜,房斌,吴磊.不均衡数据集上文本分类方法研究[J].计算机工程与应用,2013,49(20):118-121.
XIE Nana, FANG Bin, WU Lei. Study of text categorization on imbalanced data [J]. Computer engineering and applications, 2013, 49(20): 118-121.

[6] 王超学,张涛,马春森.面向不平衡数据集的改进型SMOTE算法[J].计算机科学与探索,2014,8(6):727-734.
WANG Chaoxue, ZHANG Tao, MA Chunsen. Improved SMOTE algorithm for imbalanced datasets [J]. Journal of frontiers of computer science & technology, 2014, 8(6): 727-734.

[7] 薛薇.非平衡数据集的改进SMOTE再抽样算法[J].统计研究,2012,29(6):95-98.
XUE Wei. An improved SMOTE algorithm for re-sampling imbalanced data sets [J]. Statistical research, 2012, 29(6): 95-98.

[8] 王道明,鲁昌华,蒋薇薇,等.基于粒子群算法的决策树SVM多分类方法研究[J].电子测量与仪器学报,2015,29(4):611-615.
WANG Daoming, LU Changhua, JIANG Weiwei, et al. Study on PSO-based decision-tree SVM multi-class classification method [J]. Journal of electronic measurement and instrumentation, 2015, 29(4): 611-615.

[9] 张钰莎,蒋盛益,谢柏林,等.基于改进的PSO算法的网络社区划分方法[J].计算机应用与软件,2013,30(8):25-27.
ZHANG Juesha, JIANG Shengyi, XIE Bolin, et al. Improved PSO algorithm based network community detection method [J]. Computer applications and software, 2013, 30(8): 25-27.

[10] 李晶辉,张小刚,陈华,等.一种改进隐朴素贝叶斯算法的研究[J].小型微型计算机系统,2013,34(7):1654-1658.
LI Jinghui, ZHANG Xiaogang, CHEN Hua, et al. Improved algorithm for learning hidden naive Bayes [J]. Journal of Chinese computer systems, 2013, 34(7): 1654-1658.

作者简介:高 超(1980—),男,江西南丰人,博士,副教授。主要研究方向为计算机网络、计算机软件。
许翰林(1993—),男,江苏南京人,硕士研究生。主要研究方向为计算机网络、计算机软件。

(上接第182页)

LIU Dan. Investigation of temperature control system of plastic extrusion machine [J]. China plastics industry, 2017, 45(5): 61-64.

[7] 王朋朋,黄海龙.模糊PID在粮食烘干炉温度控制系统中的应用研究[J].机械设计与制造,2017,35(2):40-42.
WANG Pengpeng, HUANG Hailong. The application study of fuzzy PID in the temperature control system of grain drying oven [J]. Machinery design & manufacture, 2017, 35(2): 40-42.

[8] 宋健.基于AVR单片机的云台控制系统设计与实现[J].现代电子技术,2016,39(13):160-162.
SONG Jian. Design and implementation of PTZ control system based on AVR microcontroller [J]. Modern electronics technique, 2016, 39(13): 160-162.

[9] 史晓娟,李松博.基于AVR单片机的嵌入式可编程控制系统[J].仪表技术与传感器,2017,20(6):54-58.
SHI Xiaojuan, LI Songbo. Embedded PLC based on AVR microcontroller [J]. Instrument technique and sensor, 2017, 20(6): 54-58.

[10] CHEN Peng, DUAN Fengyang, ZHANG Qingjie, et al. Design of UAV attitude controller based on fuzzy PID [J]. Journal of projectiles, rockets, missiles and guidance, 2015, 2(1): 9-11.

[11] WANG Weibing, ZHANG Hui, XU Qian. The design of fuzzy PID controller for temperature and pressure reducing system [J]. Journal of Harbin University of Science and Technology, 2016, 21(5): 96-100.

作者简介:王欣峰(1976—),男,山西运城人,硕士研究生,讲师。研究方向为智能控制、嵌入式控制。
任淑萍(1977—),女,山西孝义人,硕士研究生,讲师。研究方向为信息与信号处理。