基于自适应引力搜索的支持向量机在公安巡防警情分类中的应用研究

王 云 李 丛
(南京理工大学泰州科技学院 江苏 泰州 225300)

摘 要 针对传统公安警情数据手工分析结果准确性差及效率低等缺点,提出基于自适应引力搜索的支持向量机分类方法。针对 GSA(Gravitational Search Algorithm)易于陷入局部最优的特点,提出一种引力常数 G 的自适应策略,自适应调整算法的寻优步长,有效地调整了算法的探索与开发能力。将自适应 GSA 与 SVM 相结合,提出基于自适应 GSA 的 SVM 参数优化过程,利用自适应 GSA 较强的全局搜索能力,不断优化调整 SVM 参数,给出参数组合的最优解。基于自适应 GSA-SVM,与公安巡防警情数据文本分类的需求相结合,设计实现了公安警情分类系统。实验表明,改进方法优化 SVM 参数产生了较高的精度和较强的泛化能力,且针对公安巡防警情信息的分类在准确率、查全率等方面均优于传统的 GSA。

关键词 支持向量机 文本分类 公安巡防 机器学习 参数优化 引力搜索算法

中图分类号 TP311 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.07.009

APPLICATION OF SVM BASED ON ADAPTIVE GSA IN POLICE PATROL CLASSIFICATION

Wang Yun Li Cong
(*Taizhou Institute of Science and Technology*, *Nanjing University of Science and Technology*, *Taizhou 225300*, *Jiangsu*, *China*)

Abstract The manual analysis of traditional police intelligence data has some shortcomings such as poor accuracy and low efficiency of the analysis results. To solve these problems, this paper proposes SVM classification method based on adaptive gravity search. According to the characteristics of GSA that is easy to fall into local optimum, an adaptive strategy of gravitation constant G is presented to adjust the optimization step length of the algorithm adaptively. It effectively adjusts the exploration and development of the algorithm. Combining the adaptive GSA with SVM, we put forward an adaptive GSA-based SVM parameter optimization process, used the strong global search ability of adaptive GSA to continuously optimize and adjust SVM parameters, and gave the optimal solution of the parameter combination. Based on the adaptive GSA-SVM, we designed and implemented the police classification system by combining with the needs of the text classification of the police patrol alarm data. Experimental results show that the improved method has higher accuracy and stronger generalization ability in optimizing the SVM parameters. The classification of police patrol warning information is superior to the traditional GSA in terms of accuracy and recall.

Keywords SVM Text classification Public security patrol Machine learning Parameter optimization GSA

0 引 言

众多分类算法中,万普尼克(Vapnik)提出的支持向量机算法 SVM 运用尤为广泛。该算法是一种广义的线性分离器,对于特定的学习样本,无差错地进行识别或者寻找出最优的超平面[1]。SVM 在图像识别、文本分类、人脸识别、入侵检测等领域都有非常广泛的应用。

本文首先在传统 GSA 算法基础上,提出自适应引力搜索算法,将改进的算法用于对支持向量机的核函数和惩罚系数进行调优,搜索最优的 C 和 σ 的参数组

收稿日期:2019-05-20。**王云**,讲师,主研领域:数据挖掘。**李丛**,副教授。

合。最后,基于本文提出的改进方法实现了公安警情信息的高效自动分类。

1　GSA 改进

1.1　相关公式

1) 惯性质量 $M_i(t)$ 的计算公式为[2]:

$$\begin{cases} m_i(t) = \dfrac{fit_i(t) - worst(t)}{best(t) - worst(t)} \\ M_i(t) = m_i(t) / \sum_{j=1}^{N} m_j(t) \end{cases} \tag{1}$$

式中:$fit_i(t)$ 表示粒子在时刻 t 的适应度值(该值决定粒子惯性质量),$best(t)$ 代表 t 时刻最优适应度值,$worst(t)$ 代表最差适应度值。

2) 粒子 i 的引力计算公式为[3]:

$$F_{ij}^{d}(t) = G(t)\frac{M_{pi}(t) \times M_{aj}(t)}{R_{ij}(t) + \varepsilon}(x_j^d(t) - x_i^d(t)) \tag{2}$$

式中:ε 为一个非常小的常量,$G(t)$ 为万有引力常量,$R_{ij}(t)$ 表示两个粒子之间的欧式距离。

3) 万有引力常数 $G(t)$ 的一般计算公式为[4]:

$$G(t) = G_0 \times e^{-\alpha t/T} \tag{3}$$

式中:G_0 为初始值,t 为当前迭代次数,T 为总迭代次数。

4) 更新后万有引力常数 $G(t)$ 的计算公式为[5]:

$$G(t) = G_0 \times e^{-alfa\frac{t}{max\_it}} \tag{4}$$

式中:G_0 为 $G(t)$ 的初始值,$alfa$ 表示速度衰减常数,max_it 表示最大迭代次数,t 为当前迭代次数。

1.2　改进思想

由式(2)知,引力常数 G 是影响算法性能的一个参数,G 的大小直接影响算法中粒子受到的合力和加速度的大小,从而决定着算法运算时粒子每次移动"步长"的大小,是影响粒子能否摆脱局部最优、实现最优精细度最直接的因素[6]。由式(3),G 从刚开始就迅速下降,即在算法搜索初期就加大了开发力度,缺少前期的有效探索过程,容易使算法陷入局部最优。显然,这种情况容易打破算法探索和开发的平衡。

针对上述不足,本文提出了使用引力常数 G 的自适应策略来改进 GSA 算法。设计引力常量的自适应变化公式如下:

$$G'(t) = G'_0 \times \exp\left(-alfa'\left(\frac{t}{max\_it}\right)^n\right) \tag{5}$$

式中:n 为描述勘探和开发比例的一个参数,当 $n=1$ 时,式(5)将转化成式(4)。引入自适应引力常量,改进引力常量的计算公式,通过改变算法比例系数 n 来调整算法的探索与开发能力,自适应调整算法的寻优步长,使算法在初期加大搜索力度,后期加大开发力度。

1.3　性能评估

算法仿真结果如图 1 所示。图中曲线 1 为改进前引力常数随迭代次数 t 的变化情况,可见引力常量 G 从刚开始就迅速下降,易陷入局部最优;而曲线 2 则为本文改进后引力常数随迭代次数 t 的变化情况,可见提出的引力常量自适应变化公式可由 n 来调节算法的探索与开发能力,n 值越大,算法初期的搜索力度越大。

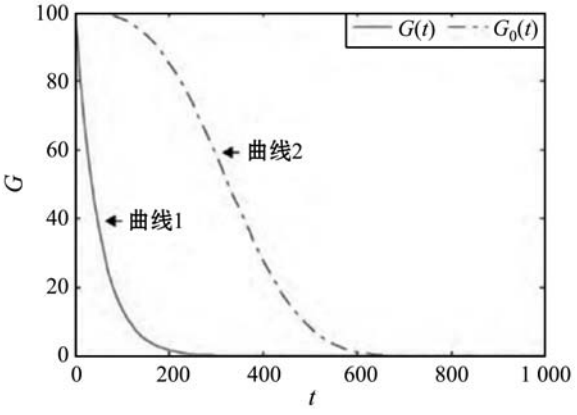

图 1　引力常量与迭代次数 t 的关系

2　基于自适应 GSA 的 SVM 参数优化

SVM 分类器性能主要取决于惩罚因子和核函数参数,而传统参数选取方法则多采用反复试凑的手工选取方法,效率低下且易获得局部最优解[7]。针对上述问题,对两个决定因素进行优化,寻找最优的参数组合显得尤为重要,下文给出基于本文提出的自适应 GSA 优化 SVM 参数的核心技术。

2.1　粒子编码方式

编码方式采用二进制字符串和 SVM 参数进行组合编码。假设一个粒子可以用一条长度为 3 的染色体来表示,其中长度为 3 表示一条染色体携带 3 个信息,分别是:进行二进制编码的字符串、惩罚系数 C 和核参数 σ。解码后对应的十进制值的计算表达式如下所示:

$$d = v_{\min} + \frac{v_{\max} - v_{\min}}{2^m - 1} \times \mathrm{dec}(\alpha) \tag{6}$$

2.2　适应度函数设计

选取一个优秀合理的适应度函数可以让实验的结

果更加准确并且具有说服力。现采用均方差(*MSE*)作为支持向量机的适应度函数[8]。*MSE* 具体公式如下:

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2 \tag{7}$$

式中:k 为训练样本数,y_i表示样本对于目标的输出,$\hat{y}_i$是对于支持向量机的输出。

2.3 SVM 参数优化流程

利用本文提出的自适应的 GSA 较强的全局搜索能力,采用自适应步长去替换原来的固定步长,不断优化调整 SVM 参数,具体算法流程图如图 2 所示。

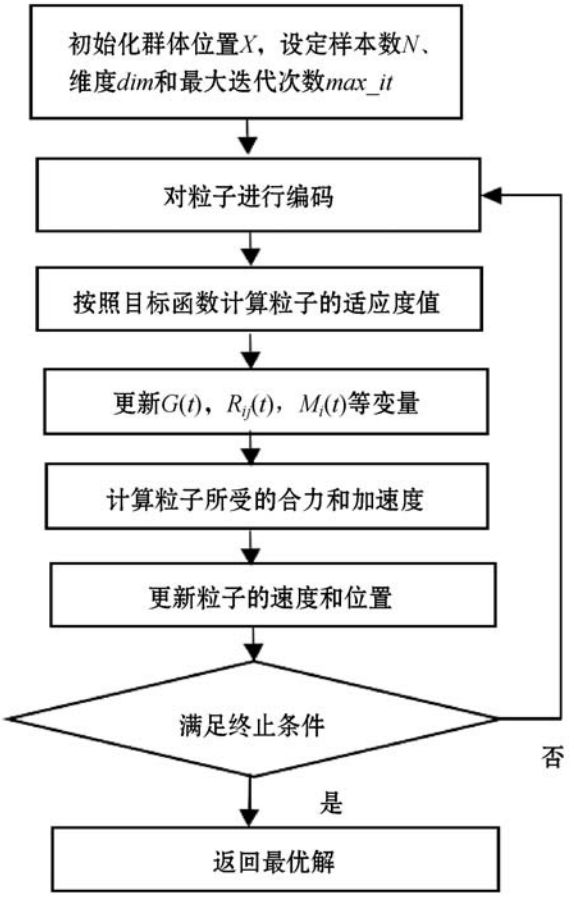

图 2 基于自适应 GSA 的 SVM 算法流程图

Step 1 随机初始化群体位置 X,设定样本数 N、维度 dim 和最大迭代次数 max_it。

Step 2 对粒子进行编码。

Step 3 计算 *MSE* 作为粒子的适应度值。

Step 4 利用式(5)计算 $G(t)$,并更新 $R_{ij}(t)$、$M_i(t)$等变量。

Step 5 计算合力、加速度。

Step 6 通过改进的自适应策略和引力计算方法计算并更新粒子的位置及速度,完成一次迭代过程。

Step 7 判断有无满足终止条件,若满足则输出算法最优解;不满足需重复算法 Step 2 – Step 6。

3 基于自适应 GSA-SVM 的公安巡防警情自动分类识别

3.1 系统功能

智能巡防系统分为基于 Android 的移动端和基于. Net 的后台管理端,移动端具备实时勤务、盘查录入、接处警、勤务查询、在线学习、警情自动分类等功能;后台具有对信息的管理功能,主要包括删改查等数据操作。其中警情自动分类系统是泰州市海陵区公安巡防智能系统的子系统,其分类需求主要包括交通事故类别、反恐类别、刑事类别等。该子系统可以实现将采集的情报文本自动归类到已设定的类别中,便于案情研判者选择其感兴趣的类别,并与同类别或不同类别信息进行对比分析。

3.2 开发环境

本系统基于 Windows 10 操作系统,使用 Java 编程语言编制智能巡防分类模块核心代码,并将其打包成一个功能 jar 包,供外部应用调用。

3.3 多分类器构建

公安巡防信息中,警情类别有多种,针对二分类问题的 SVM 算法则显得无能为力。本文采用间接法构造多个分类器,从而克服传统 SVM 算法分类的不足,根据需要,共需 $n(n-1)/2$ 个 SVM,而其中每个 SVM 均采用二分类训练集进行训练[9]。例如,以下给出在 a 和 b 两个类中寻找最优的超平面训练集:

$$\begin{cases}(x_{ab,t}, y_{ab,t}) \\ x_{ab,t} \in \mathrm{R}^d, y_{ab,t} \in \{a, b\}\end{cases} \quad t = 1, 2, \cdots, n_{ab} \tag{8}$$

$$\min_{\omega_{ab}, \varepsilon_{ab}} \frac{1}{2}\|\omega_{ab}\|^2 + C\sum_{a=1}^{m_{ab}}\xi_{ab,t} \tag{9}$$

$$\text{s.t.}\begin{cases}(\omega_{ab})^{\mathrm{T}}\varphi(x_{ab,t}) + b_{ab} \geqslant 1 - \xi_{ab,t}, y_{ab,t} = a \\ (\omega_{ab})^{\mathrm{T}}\varphi(x_{ab,t}) + b_{ab} \geqslant -1 + \xi_{ab,t}, y_{ab,t} = b \\ \xi_{ab,t} \geqslant 0\end{cases} \tag{10}$$

在建立 $n(n-1)/2$ 个 SVM 模型基础上,对检测样本进行判断分类。经过筛选淘汰之后,最终输出的类别就是测试样本类别。

3.4 警情文本分类模块设计

警情自动分类模块由训练样本和分类两部分组成。对于已知类别的训练样本,采用分类算法计算分类模型;对于测试样本,使用上步计算结果的分类模型判断待分类样本点所属类别,后对分类器的性能进行评价。警情文本分类模块设计如图 3 所示。

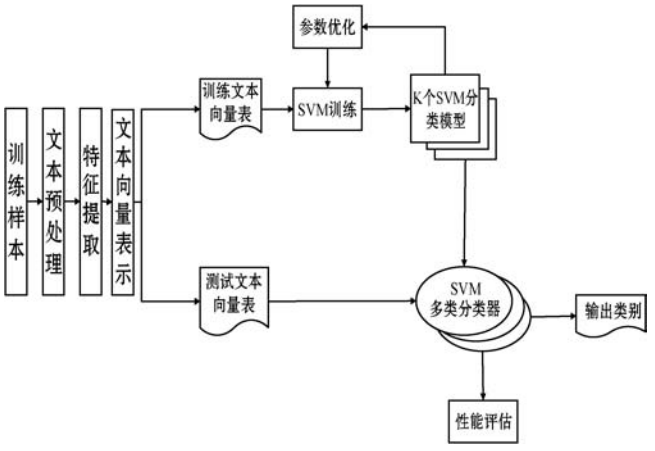

图 3　警情文本分类模型设计图

3.5　警情分类识别的步骤

本文所实施的警情分类识别步骤如下:

Step 1　收集原始的文本数据。

Step 2　对数据进行预处理,使得处理后的结果可以作为自动分类子系统的输入,减少样本训练使用的时间,加快收敛速度,从而加强对警情的判别能力。

Step 3　把预处理的数据交由 SVM 作为输入,通过自适应 GSA-SVM 算法优化得到最优参数。

Step 4　使用 Step 3 结果进行样本训练时,针对不同数量的警情类别样本数,须采用不同的方法训练样本(样本数较多时,采用边界样本方法,否则采用虚拟样本),最后进行组合构造,最终建立最优警情判别模型。

Step 5　基于 Step 4 建立的警情判断模型,加入测试样本集合进行检测。

Step 6　输出加入的测试样本的警情判断结果。

4　实验评估

为了验证算法改进的效果及公安巡防自动分类系统设计的合理性,本文测试选取警情中的多个类别,多组数据综合进行验证。

4.1　测试数据与评估标准

1) 测试数据。文本分类中的语料库指用于测试和训练学习机器的文本集合,语料库选择是否合适,将直接影响文本分类器的性能。语料库应广泛代表分类系统所需处理的实际存在的各类别文本。本文测试中,语料库来源于泰州市海陵区公安情报文本集,训练文本共计 1 521 条,测试文本共计 612 个。测试分为 10 组进行,为了前后测试的连贯性,每组测试中训练文本集始终保持不变。测试用数据分布要求如表 1 所示。

表 1　测试用的数据分布表

编号	类别	训练文本数	测试文本数	合计
1	治安	252	88	340
2	刑事	346	184	530
3	交通	325	64	389
4	火灾	123	75	198
5	救助	180	69	249
6	缉毒	152	61	213
7	反恐	143	71	214
合计		1 521	612	2 133

2) 性能评估标准。采用目前普遍使用的准确率和查全率以及两者的综合评价指标 F1 测试值作为分类器性能的评价指标[10]。

4.2　实验结果与分析

1) SVM 参数优化方法的有效性测试。首先,确定自适应引力搜索算法的初始参数,如:样本数为 1 200,维数为 800,最大迭代次数为 30。然后在编码范围里随机寻找多个点作为粒子群的初始位置,按照适应度函数来搜索全局最优点。详细的实验结果如表 2 所示。

表 2　SVM 参数优化结果

迭代次数	惩罚因子	核函数参数	适应度/%	错分率/%
1	1 055.568	201.526	67.526	32.526
4	855.326	52.326	69.523	30.526
8	660.218	102.362	71.523	25.913
10	752.329	65.361	74.856	21.206
14	552.142	35.946	78.523	20.412
16	875.957	40.523	79.126	17.562
20	752.580	25.624	81.543	15.256
23	946.625	17.529	82.523	14.236
25	936.597	14.526	83.453	13.526
27	935.659	12.362	84.579	12.203
29	937.591	11.842	85.419	12.175
30	935.246	11.842	86.947	12.006

表 2 列出了在 SVM 参数的优化过程中,随着迭代次数的增加,得到了 (C, σ^2) 的最优解(935.246, 11.842)。通过查看表中数据,可以发现随着迭代次数的增加,错分率在下降的同时,适应度有所提高。

2) 分类性能比较。为了便于比较,现分别采用 GSA 和本文改进的自适应 GSA 算法对 SVM 进行训练,最大迭代次数设为2 000,粒子数为 40,G_0 设为 50。采用泰州市海陵区公安巡防智能系统情报文本集中的 100 个警情文本数据作为测试集,两种方法比较结果

如表 3 所示。

表 3 性能比较

SVM 参数方法	训练 MSE 练 MSE	测试 MSE
GSA	0.022 1	0.243
改进 GSA	0.015 9	0.128

结果表明,基于自适应的 GSA 优化 SVM 参数产生了较高的精度和较强的泛化能力。

两种方法对公安巡防警情信息的分类效果比较如图 4 所示。

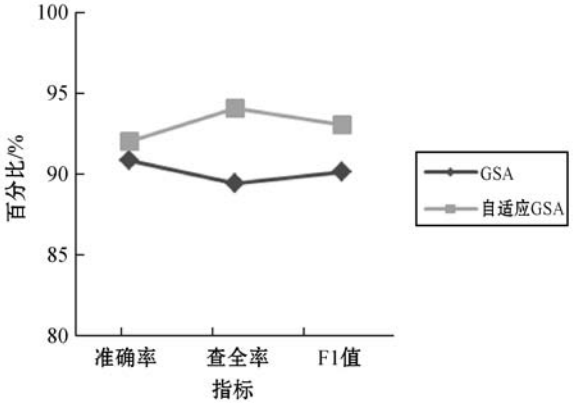

图 4 两种方法比较 SVM 分类的效果

可以看出,本文所用的 SVM 参数优化方法无论是在准确率、查全率及综合评价指标 F1 方面,都优于传统 GSA 算法,采用基于自适应 GSA 的 SVM 方法可以得到更优的 SVM 参数。

5 结 语

本文在传统 GSA 的基础上,提出一种自适应的引力搜索算法。该算法避免了传统 GSA 易于得到局部最优解的缺点,用改进的 GSA 优化 SVM 参数,参数优化结果的精度有所提高。将研究的方法应用于实际的公安巡防警情自动分类中,实验结果表明,本文提出的改进方法是合理、有效的。

参 考 文 献

[1] 贝雨馨,崔荣一. 文本分类中特征项权重的计算方法[J]. 延边大学学报(自然科学版),2004,30(3):202 - 204.

[2] Coelho F, Braga A P, Verleysen M. A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems[J]. International Journal of Computational Intelligence Systems, 2016, 9(4): 726 - 733.

[3] Rashedi E, Nezamabadi-Pour H, Saryazdi S. GSA: A gravitational search algorithm[J]. Intelligent Information Management, 2012, 4(6): 390 - 395.

[4] 王蕾,潘丰. 改进多样性和局部优化能力的引力搜索算法[J]. 计算机工程,2014,40(8):173 - 178.

[5] 马占飞,陈虎年,杨晋,等. 一种基于 IPSOSVM 算法的网络入侵检测方法[J]. 计算机科学,2018,45(2):231 - 235,260.

[6] 敖媛,丁学明. 基于自适应引力搜索算法的 T-S 模型辨识[J]. 系统仿真学报,2017,29(3):487 - 493.

[7] 牛琳. 基于 SVM 的公安情报自动分类系统的设计与实现[D]. 郑州:中国人民解放军信息工程大学,2007.

[8] 杨晋,金溢,马占飞. 基于 IQPSO 算法的网络入侵检测研究[J]. 内蒙古科技大学学报,2018,37(1):96 - 102.

[9] 董新燕,丁学明,王健. 基于改进的引力搜索算法的 T-S 模型辨识[J]. 电子科技,2015,28(11):16 - 20.

[10] 陶俐言,杨海斌. 基于改进引力搜索算法的公差多目标优化设计[J]. 机械设计与研究,2017,33(2):133 - 137.

(上接第 55 页)

[5] 陈奎,陈博博. 基于改进暂态相关分析和支持向量机的电弧故障选线研究[J]. 电力系统保护与控制,2016,44(24):66 - 73.

[6] 陈奎,韦晓广,陈景波,等. 基于样本数据处理和 ADABOOST 的小电流接地故障选线[J]. 中国电机工程学报,2014(34):6228 - 6237.

[7] Guo M F, Zeng X D, Chen D Y, et al. Deep-learning-based earth fault detection using continuous wavelet transform and convolutional neural network in resonant grounding distribution systems[J]. IEEE Sensors Journal, 2018, 18(3): 1291 - 1300.

[8] Luo G M, Yao C Y, Liu Y L, et al. Stacked auto-encoder based fault location in VSC-HVDC[J]. IEEE Access, 2018, 6: 33216 - 33224.

[9] Dai J J, Song H, Sheng G H, et al. Cleaning method for status monitoring data of power equipment based on stacked denoising autoencoders[J]. IEEE Access, 2017, 5: 22863 - 22870.

[10] Chen K J, Hu J, He J L. A framework for automatically extracting overvoltage features based on sparse autoencoder [J]. IEEE Transactions on Smart Grid, 2018, 9(2): 594 - 604.

[11] 葛眠俊. 基于脉冲神经膜系统的小电流单相接地故障选线[D]. 四川:西南交通大学,2018.

[12] 束洪春,黄海燕,田鑫萃,等. 采用形态学峰谷检测的谐振接地系统故障选线方法[J]. 电力系统自动化,2019,43(1):228 - 233.

[13] Cheng G, Yang C, Yao X, et al. When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs[J]. IEEE Transactions on Geoscience & Remote Sensing, 2018, 56(5): 2811 - 2821.

[14] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders[C]// 25th International Conference on Machine Learning, 2008: 1096 - 1103.