基于云模型的并行蚁群-SVM 分类方法

余桂兰,陈 珂,左敬龙
(广东石油化工学院 计算机与电子信息学院,广东 茂名 525000)

摘 要:支持向量机(SVM)是一种高效的分类识别方法,在解决高维模式识别问题中表现出许多特有的优势,但 SVM 不利于海量数据的挖掘。为了改善 SVM 对大样本数据的适应性,提高算法的收敛速度,利用云模型来优化并行蚁群算法,提出了一种基于云模型的并行蚁群-SVM 网页分类方法。将蚂蚁当前位置坐标作为云滴的两个参数,用逆向云发生器产生信息云的三个数字特征,采用不同的方法来更新蚂蚁的信息素,比较真实地体现了现实蚁群的运作情况,达到了实时动态更新的效果。通过对比测试,验证了 CPACA-SVM 方法在准确率和召回率上均有明显提高,具有较好的分类效果。

关键词:云模型;逆向云发生器;并行蚁群算法;支持向量机;网页分类
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2014)04-0131-04
doi:10.3969/j.issn.1673-629X.2014.04.033

Parallel Ant Colony-SVM Classification Method Based on Cloud Model

YU Gui-lan, CHEN Ke, ZUO Jing-long
(College of Computer and Electronic Information, Guangdong University of Petrochemical Technology, Maoming 525000, China)

Abstract: SVM is an effective method for learning the classification knowledge from massive data, especially in solving the high dimensional pattern recognition problem. But SVM is not conducive to massive data mining. To improve the adaptability of SVM on large-scale scenes and the speed of convergence of the algorithm, utilizing the cloud model to optimize parallel ant colony algorithm, a parallel ant colony-SVM web page classification method based on cloud model is proposed. Three digital characteristics of the cloud is produced from backward cloud generator, the ants current position coordinates is composed of the two parameters of cloud droplets. Using different methods to update the ant pheromones, more accurately reflect the life of the ant colony, to achieve the effect of real-time dynamic updates. By comparison test verify the CPACA-SVM method on precision and recall rate significantly improved, with better classification effect.

Key words: cloud model; backward cloud generator; parallel ant colony algorithm; SVM; web page classification

0 引 言

随着网络应用及计算机的普及,人们对网络的需求越来越大,以至于网站数目及其包含的网页数目均呈爆炸式增长。面对 Web 上海量的文档,如何快速高效地对各式各样的文档进行准确的分类,是当前信息处理研究领域的热点问题之一。目前采用的主要方法有:最近邻(K Nearest Neighbor, KNN)分类法[1]、决策树(Decision Tree)分类法[2]、Bayes 分类法[3]、支持向量机(Support Vector Machines, SVM)分类法[4]以及各种混合算法等。

支持向量机是一种建立在统计学习理论基础上的可训练机器学习方法。它根据小样本学习后的模型参数进行特征提取,可以得到分布均匀的类表。SVM 可以自动寻找那些对分类有较好区分能力的支持向量,由此构造出的分类器可以最大化类与类的间隔,因而有较好的推广性能和较高的分类准确率。SVM 的出色表现,使得其应用广泛,但同时,SVM 也存在对大规模数据适应性不强、收敛速度慢等缺陷。为了改善这些问题,文献[5]将支持向量机和蚁群算法相结合来构造网页分类器;文献[6]则提出了将具有量子特性的 ACA 和 SVM 进行融合的网页分类方法,采用动态调整旋转角的策略来及时更新信息素;文献[7]将支持向量机与 K 近邻进行结合来完成网页分类;文献[8]提出了一种基于紧密度的模糊支持向量机方法,

收稿日期:2013-06-19 **修回日期**:2013-09-26 **网络出版时间**:2014-01-28
基金项目:广东省科技计划项目(2012B010100037);茂名市科技计划项目(2012B01040)
作者简介:余桂兰(1975-),女,硕士,CCF 会员,研究方向为计算智能、网络优化、云计算;陈 珂,副教授,硕士,研究方向为数据挖掘、机器学习;左敬龙,副教授,硕士,高级程序员,研究方向为网络安全、虚拟技术、云计算。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140128.1201.060.html

在确定样本的隶属度时,考虑了样本与类中心、样本之间的关系来提高隶属度函数的适应性。这些算法在某种程度上取得了一定的效果,但大都存在主观性,如隶属度函数采用人为选取,存在一定的局限性。鉴于此,文中提出了一种基于云模型的并行蚁群算法与支持向量机结合的网页分类方法,该方法引用了李德毅院士提出的云理论,规避了很多算法在采用模糊技术处理时人为选定隶属度函数这一关键步骤,比较客观、准确地反映出系统中样本存在的不确定性,更好地改善了 SVM 对大规模数据的适应性问题,使得算法在收敛速度上有较大提高。

1 云模型与支持向量机

1.1 云模型

云模型[9-10](Cloud Model)通过语言值的概念表示,把不确定概念的随机性和模糊性有机地结合在一起,利用 3 个数字特征(期望值 Ex、熵 En、超熵 He)刻画了自然科学与社会科学中存在的大量模糊、不确定性现象,实现了不确定语言值与定量数值之间的自然转化。设 U 是一定量论域,C 是 U 上的定性概念,(Ex,En,He)为 C 的数字特征。若定量值 $x \in U$ 且 x 是对于 C 的确定度为 $y=\mu C(x)$ 的一次具有稳定倾向的随机实现,则云滴(x,y) 在论域 U 上的分布称为云。

若 x 满足 $x \sim N(\mathrm{Ex},\mathrm{En}'^2)$, 其中 $\mathrm{En}' \sim N(\mathrm{En},\mathrm{He}^2)$ 是期望为 En、方差为 He^2的正态分布,且

$$y = \mathrm{e}^{-\frac{(x-\mathrm{Ex})^2}{2\mathrm{En}'^2}}$$

则 x 在论域 U 上的分布 X 称为正态云,其概率密度为:

$$f_x(x)=\int_{-\infty}^{\infty}\frac{1}{2\pi\mathrm{He}\,|y|}\mathrm{e}^{-\frac{(x-\mathrm{Ex})^2}{2y^2}-\frac{(y-\mathrm{En})^2}{2\mathrm{He}^2}}\mathrm{d}y$$

云的数字特征具体定义为:期望 Ex 表示云滴在论域空间分布的期望,它反映了相应模糊概念的信息中心值;熵 En 表示定性概念的粒度,即可接受的数值范围;超熵 He 是熵 En 的不确定性度量,它反映了云滴的离散程度, 揭示了模糊性和随机性的关联。

逆向云发生器如图 1 所示。

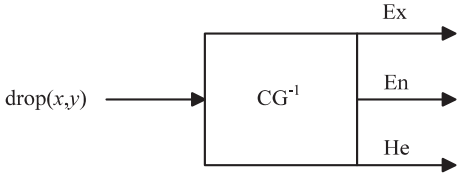

图 1 逆向云发生器

1.2 支持向量机及其训练

支持向量机是从线性可分情况下的最优分类面发展而来的。通过将向量映射到一个更高维的空间 S,在空间 S 里建立一个最大间隔的分类超平面 H,利用核方法先做一个从原空间到核特征空间的映射 $x \rightarrow \phi(x)$,然后在特征空间构造超平面 $(W \bullet \phi(x))+b=0$。这样,只要映射 ϕ 构造恰当,就可将原空间的线性不可分问题转化为核特征空间的线性可分问题,而 ϕ 的构造可以采用构造核特征空间的内积 $\phi(x)\bullet\phi(y)=K(x,y)$ 来间接完成(文中采用的是满足 Mercer 核的高斯核函数[11])。最优分类面就是在 H 的两边建立两个互相平行于 H 的超平面 H_1、H_2,使得它们之间的间隔(即分类间隔 margin)最大化,且能将两类样本正确分开。从而将问题转化为在约束条件下求解二次规划问题(QP)。SVM 分类示意图如图 2 所示。

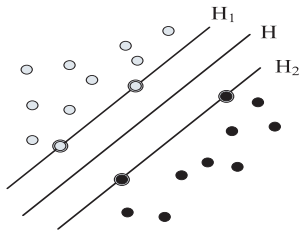

图 2 SVM 分类示意图

采用 TF-IDF 特征表示法对网页进行预处理,将网页表示成由词条组成的向量形式,然后经过特征提取、降维后,开始 SVM 训练。

训练一个 SVM 也就相当于求解下述 QP 问题:假设训练集 $E=\{(x_i,y_i)\mid i=1,2,\cdots,n\}$,其中 $x_i \in R^n$,$y_i \in (-1,+1)$,并假设参数 ρ 解决下面的二次优化问题:

$$\min: W(\rho)=-\sum_{i=1}^{n}\rho_i+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho_i\rho_j y_i y_j K(x_i\cdot x_j) \tag{1}$$

$$\mathrm{sub}:\ \sum_{i=1}^{n}\rho_i y_i=0$$

$\forall i: 0 \leqslant \rho_i \leqslant C$,C 是折衷参数,控制数据在训练样本中的影响。

求解上述问题,得到一个最优的判决函数:

$$f(x)=\mathrm{sgn}\Big[\sum_{i=1}^{n}\rho_i y_i(x\bullet x_i)+b\Big]$$

其中,ρ_i为 Lagrange 乘子;b 为分类面的阈值。

结合非线性可分的情况,将 x 做变换 $\Phi: R^d \rightarrow H: x \rightarrow \Phi(x)=(\phi_1(x),\phi_2(x),\cdots,\phi_i(x),\cdots)^{\mathrm{T}}$, d 为空间维数,$\phi_i(x)$为实函数。于是有:

$$f(x)=\mathrm{sgn}\Big[\sum_{i=1}^{n}\rho_i y_i(\Phi(x)\bullet\Phi(x_i))+b\Big]=$$

$$\mathrm{sgn}\Big[\sum_{i=1}^{n}\rho_i y_i K(x,x_i)+b\Big]$$

再将 $K(x,x_i)$替换成高斯基 RBF 核函数,$K(x,x_i)=\exp(-\|x-x_i\|^2/(2\sigma)^2))$,$\sigma>0$。于是可以得到最优解:

$$\rho^*=(\rho_1^*,\rho_2^*,\cdots,\rho_i^*,\cdots,\rho_l^*) \tag{2}$$

这些 ρ_i 中,只有一部分位于分界线附近的系数 ρ_i 非零,它们对应的样本向量就是支持向量,只有支持向量影响最终的划分结果,其他都可以看作冗余向量。

2　基于云模型的改进型并行蚁群-SVM 网页分类方法

2.1　基于云模型的并行蚁群算法

云模型并行蚁群算法(Cloud Parallel Ant Colony Algorithm,CPACA)是将云模型与并行蚁群算法[12]进行有机融合,将蚂蚁当前位置坐标作为云滴 drop(x, y)的两个参数,用逆向云发生器产生信息云的三个数字特征(Ex,En,He),采用局部和全局两种方法来更新蚂蚁的信息素,以求达到完全实时动态更新的效果。

2.1.1　并行蚁群算法

并行蚁群算法是将一个总的蚁群分成若干个子蚁群,不同的蚁群赋予不同的控制参数,各蚁群相互独立寻优,在独立运行若干代后,利用群体间的交互作用,实现多个蚁群的协同进化,从而得到最优解。它真实地体现了蚂蚁社会的实际运作和蚁群算法的分布性特征,在解决复杂的优化问题上具有比较优越的性能,但也存在进化速度慢、容易陷入局部最优的情况。因此,利用云模型在知识表示中的特点,将云模型和并行蚁群算法进行结合,使算法在定性知识的指导下能够自适应控制搜索空间范围,能在较大搜索空间条件下避开局部最优解。

假设将所有蚂蚁划分为 k 个种群,每个种群有 n 只蚂蚁,同一种群内的蚂蚁释放同一种信息素,不同种群释放的信息素不同,种群内的信息素对蚂蚁有吸引作用,种群间的信息素对蚂蚁有排斥作用。则第 k 种群的蚂蚁在时间 t 从样本点 i 以概率 p(定义见式(3))在其领域 Π_l 内选择下一样本点 j 的吸引因子为:

$$\alpha_{ij}^k = \frac{\tau_{ij}^k}{\sum_{h\in\Pi_i}(\tau_{ih}^k)}$$

排斥因子为:

$$\beta_{ij}^k = \frac{\sum_{h\neq k}(\tau_{ij}^k)}{\sum_{h\in\Pi_i}(\tau_{ih}^k)}$$

其中,τ_{ij}^k 为路径(i,j)上的信息素浓度。

吸引因子 α_{ij}^k 越大,说明路径(i,j)上的信息素对 k 群内蚂蚁的吸引力越大,排斥因子 β_{ij}^k 越大,说明路径(i,j)上的其他种群信息素对 k 群蚂蚁的排斥力越大。

$$p_{ij}^k(t) = \begin{cases} \dfrac{[\tau_{ij}^k(t)^{\frac{\varepsilon\alpha_{ij}^k}{\beta_{ij}^k}}]\bullet[\eta(t)^{\beta_{ij}^k}]}{\sum_{k\notin \mathrm{tabu}_k}[\tau_{ij}^k(t)^{\frac{\varepsilon\alpha_{ij}^k}{\beta_{ij}^k}}]\bullet[\eta(t)^{\beta_{ij}^k}]} & j\notin \mathrm{tabu}_k \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

其中,tabu_k 是禁忌列表,每个节点维护一张信息素表,记录邻边上信息素的浓度;$\varepsilon>0$ 为调整因子,用以调节信息素信息的影响权重。

2.1.2　信息素更新策略

算法将云模型应用到信息素更新策略中,将信息素更新分为局部更新和全局更新。采用云模型将 k 种群中所有蚂蚁的同一种信息素的分布特征定义为信息云,k 种群的第 i 种信息素的信息云记为 $C=(\mathrm{Ex}_i,\mathrm{En}_i,\mathrm{He}_i)$,则局部信息素更新规则定义为:

$$\tau_{ij}^k = \mathrm{En}_i(\mu_i - x_i) + \mathrm{Ex}_i(\mu - x_i) + \mathrm{He}_i(\mu_j - x_i) \quad (4)$$

其中,En_i、Ex_i 和 He_i 为第 i 种信息素的信息云的三个特征值,具体做法是将蚂蚁当前位置坐标 x,y 代入云滴 drop(x,y)中,用逆向云发生器产生信息云的三个数字特征(Ex,En,He);μ 为全局极值;μ_i 为个体所在种群极值;μ_j 为领域种群极值。

全局信息素更新规则为:

$$\tau_{ij}^k = \tau_{ij}^k + (\rho^k)^{-1}, (\rho^k \text{为式(2)中的第 } k \text{ 种群 } \rho^*) \quad (5)$$

由于 $C=(\mathrm{Ex}_i,\mathrm{En}_i,\mathrm{He}_i)$ 中的三个参数是由蚂蚁当前位置信息转换而来的,因此信息素更新策略具有实时动态性质,完全规避了主观经验,是一种比较客观、准确的更新方法。

2.2　CPACA-SVM 算法

算法优化求解开始时,训练集中的每一个样本点对应于一个蚂蚁智能体,从 $t=0$ 时刻开始进行第一轮搜索,禁忌列表记录了每只蚂蚁所经过的样本点路径,当蚂蚁遍历了所有样本点时,便完成一次循环,此时蚂蚁走过的路径便是问题的一个侯选解。

算法过程描述如下:

(1)初始化:设置最大次数 Max,清空禁忌列表 tabu(l),初始信息素 τ_{li}^0 为某一相同值,$t=1$(t 为迭代次数);

(2)将所有蚂蚁划分为 k 个种群,每个种群有 n 只蚂蚁,将它们随机分配到 m 个训练集中;

(3)当未达到最大迭代次数时,执行(4)~(13);

(4)对每个子群并行执行(5)~(8);

(5)当第 l 个个体经过第 j 个子区间,如果样本点 j 不在禁忌列表 tabu(l)中,则执行(5)~(6);

(6)根据式(3)选择规则,计算个体 l 在样本点 i 选择样本点 j 作为下一站的概率,并按概率选择样本点;将个体 l 移至该样本点,并将编号放入个体 l 的禁忌列表(式(3))中;

(7)计算该个体经过的路径长度,记录该个体所在种群极值 μ_i 和领域种群极值 μ_j;

(8)将蚂蚁当前位置坐标作为云滴 drop(x,y)的两个参数,用逆向云发生器产生信息云的三个数字特

征(Ex,En,He),运用式(4)对信息素进行更新;

(9)记录本轮搜索到的最优解及全局极值μ;

(10)清空禁忌列表;

(11)应用式(1)对本轮得到的最优解进行一轮训练;

(12)$t=t+1$;

(13)根据式(5)对所有路径(i,j)更新信息素;

(14)输出最优解;将最优解交给 SVM 进行训练,最终获得 SVM 分类器;对网页进行分类。

3 实验及分析

实验数据来自网络下载的 3 000 中文网页,共八大类别。任意抽取 2/3 网页作为训练集,剩下的作为测试集。评测标准采用通用的准确率(Precision)、召回率(Recall)和测试值 F1[13]。首先进行网页预处理,然后将文中算法 CPACA-SVM 与文献[5]的 ACA-SVM、文献[6]的 QACA-SVM 方法分别进行比较。实验分两轮进行:首先进行单一测试,分别测试八大类别的准确率、召回率,测试结果见图 3;然后进行总体测试,不分类别,结果如图 4 所示。可以看出:在分类测试中,由于数据量相对较小,三种算法的准确率相差不大,在召回率上 CPACA 表现较好,从整体看,CPACA 最优;图 4 则充分反映出 CPACA 的优越性,三个指标

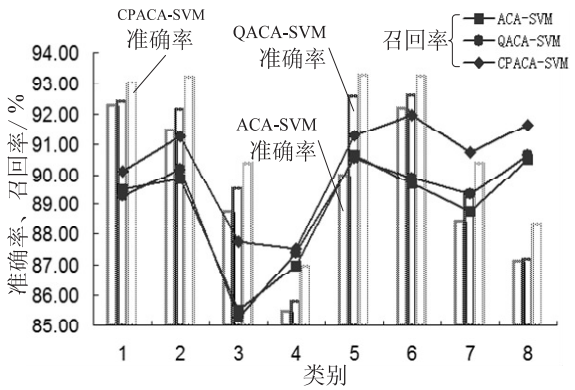

图 3 八大类别单一测试

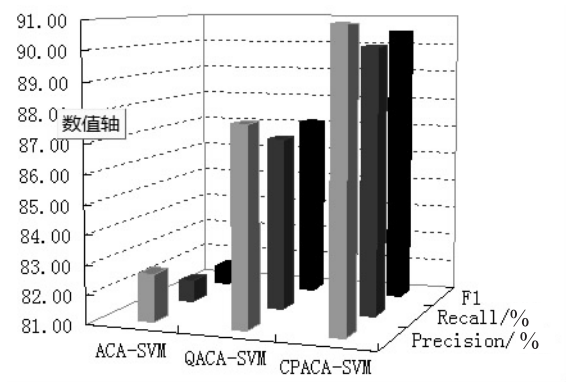

图 4 总体测试

都表明了该基于云模型的并行蚁群-SVM 网页分类方法具有较好的分类效果。

4 结束语

分类是 Web 挖掘研究的一个关键步骤,网页分类由于具有半结构化、格式多变、内容繁多、来源复杂、增长迅速等特点,一直是一个很重要的研究课题。文中以支持向量机为基础,利用云模型来优化并行蚁群算法,提出了一种基于云模型的并行蚁群-支持向量机的网页分类方法。由于该方法在更新蚂蚁信息素时,采用的是实时自动更新策略,不存在人为指定因素,故该算法具有迭代次数少、运行时间短及分类效果好等优点。通过与其他算法进行对比测试,CPACA-SVM 方法在准确率和召回率上均优于其他算法,在一定程度上提高了 SVM 的分类效果。

参考文献:

[1] 闫 晨. KNN 文本分类研究[D]. 秦皇岛:燕山大学,2010.

[2] He Jieyue. Rule generation for protein secondary structure prediction with support vector machines and decision tree[J]. IEEE transactions on nanobioscience,2006,5(1):46-53.

[3] Yan Zhiyong, Xu Congfu, Pan Yunhe. Improving naive Bayes classifier by dividing its decision regions[J]. Journal of Zhejiang University-Science C(Computers & Electronics),2011,12(8):647-657.

[4] 梁 燕. SVM 分类器的扩展及其应用研究[D]. 长沙:湖南大学,2008.

[5] 宋军涛,杜庆灵. 基于改进蚁群算法和支持向量机的网页分类研究[J]. 电脑知识与技术,2009,5(35):10069-10071.

[6] 左敬龙,余桂兰. 具有量子特性的 ACA-SVM 网页分类方法[J]. 计算机工程与应用,2011,47(12):49-51.

[7] 宗永升,张 祎. 支持向量机与 K 近邻结合的网页分类方法[J]. 计算机仿真,2010,27(9):208-211.

[8] 张 翔,肖小玲,徐光祐. 基于样本之间紧密度的模糊支持向量机方法[J]. 软件学报,2006,17(5):951-958.

[9] 刘 禹,李德毅,张光卫,等. 云模型雾化特性及在进化算法中的应用[J]. 电子学报,2009,37(8):1651-1658.

[10] 付 斌,李道国,王慕快. 云模型研究的回顾与展望[J]. 计算机应用研究,2011,28(2):420-426.

[11] 赵 莹. 支持向量机高斯核函数的研究[D]. 上海:华东师范大学,2007.

[12] 陈 峻,章春芳. 并行蚁群算法中的自适应交流策略[J]. 软件学报,2007,18(3):617-624.

[13] 张启蕊,董守斌,张 凌. 文本分类的性能评估指标[J]. 广西师范大学学报:自然科学版,2007,25(2):119-122.

基于云模型的并行蚁群-SVM分类方法

作者： 余桂兰， 陈珂， 左敬龙， YU Gui-lan， CHEN Ke， ZUO Jing-long
作者单位： 广东石油化工学院 计算机与电子信息学院,广东 茂名,525000
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(4)