

代价与样本相关的简约核支持向量机

何海江

(长沙学院 计算机教学中心, 长沙 410003)

(haijianghe@sohu.com)

摘要:针对机器学习领域中误分类代价与样本相关的情况,提出一种以最小化总代价为目标的样本相关代价敏感的简约核支持向量机 sd2sSVM。首先,在 GSVM 框架下,将优化目标转换为无约束数学规划问题,再引入分段多项式平滑函数逼近正号函数,使用 Newton-YUAN 方法求无约束问题的唯一最优解,最后引入简约核提高解非线性问题的效率。实验结果表明,与传统的样本相关代价敏感支持向量机相比,sd2sSVM 的分类精度、误分类代价相当,但训练时间、预测时间则更短。另外,讨论了参数 C 对 sd2sSVM 分类性能的影响。

关键词:代价敏感;简约核;无约束;支持向量机;分类

中图分类号: TP18 **文献标志码:** A

Costs sensitive to examples learning in reduced support vector machine

HE Hai-jiang

(Computer Teaching Center, Changsha University, Changsha Hunan 410003, China)

Abstract: In many machine learning domains, misclassification costs are sensitive to examples. As an extension of class dependent costs, a cost-sensitive reduced Support Vector Machine (SVM) named sd2sSVM that aimed at minimizing all costs was introduced. Firstly, through the use of Generalized SVM (GSVM) framework, the optimization object was converted into unconstrained mathematical programming problems. Secondly, based on smooth piecewise polynomial function that was used to approach the plus function, the unique optimization solution can thus be gained by Newton-YUAN method. Finally, reduced kernel was employed to improve the solution of nonlinear problem. The experimental results show that sd2sSVM is comparable or choicer than traditional example dependent cost-sensitive SVM. It was also discussed that how parameter C influenced the performance of sd2sSVM.

Key words: cost-sensitive; reduced kernel; unconstrained; Support Vector Machine (SVM); classification

0 引言

分类是机器学习领域最重要的任务,分类算法通常假定所有类别的误分类代价相同,强调分类的准确率。但在实际应用中,不同类别的误分类往往导致不同的损失,代价敏感的分类算法成为解决问题的关键。支持向量机(Support Vector Machine, SVM)以结构风险最小化为优化目标,与贝叶斯、KNN(K Nearest Neighbor)、决策树、神经网络等其他机器学习算法相比,SVM 具有良好的泛化能力。SVM 在文本自动摘要、网页或文本自动分类、函数回归及逼近、非线性控制^[1]等问题中,都有非常成功的应用。因此,代价敏感 SVM^[2-3](cost-sensitive SVM, csSVM)成为新的研究热点,分类器的设计目标由原来的最小化分类精度变为最小化总体误分代价。

然而,现实世界有一些分类问题,同一类别的样本错误分类损失并不相同^[4],样本的误分代价依赖于样本而非类别。自然地,代价与样本相关的 csSVM 才能解决此类问题。文献[5]和文献[6](源代码可从 <http://www.cs.cornell.edu/People/tj/svm%5Flight> 下载)提出和实现了一次松弛(式(1)的 $k=1$)的代价样本相关 csSVM,能够降低总的误分代价,但支持向量数目过大。受到平滑 SVM^[7]和简约核^[8]成功应用的启发,我们提出一种代价与样本相关的二次松弛平滑型简约核 SVM (Costs Sensitive to Examples 2-Norm Smooth Reduced SVM, sd2sSVM),支持二次松弛(式(1)的 $k=2$)的代价样本相关 csSVM,与前者的分类精度、误分类代价相当,

但训练时间更短,预测待分类样本的类别也更快,能应用于实时性要求高的分类系统。

1 代价与样本相关的 sd2sSVM

我们只关注两类问题的分类,多类问题可转化为两类问题。假设训练样本集 $S = \{(\mathbf{x}_i, y_i) \mid i = 1, \dots, m\}$, 输入向量 $\mathbf{x}_i \in R^n$ 是列向量,目标值 $y_i = \{+1, -1\}$ 。 I_{+1} 和 I_{-1} 分别是正例和负例样本的索引集,即: $I_{+1} = \{i \mid y_i = +1\}$, $I_{-1} = \{i \mid y_i = -1\}$, 而 $c_{+1}(\cdot)$ 和 $c_{-1}(\cdot)$ 是正例和负例的误分代价函数。结构风险最小化分类问题可形式化为以下优化问题:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i \in I_{+1}} c_{+1}(\mathbf{x}_i) \xi_i^k + C \sum_{i \in I_{-1}} c_{-1}(\mathbf{x}_i) \xi_i^k \quad (1)$$

约束条件为:

$$\begin{cases} y_i (\mathbf{w}^T \Phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \\ i = 1, 2, \dots, m \end{cases} \quad (2)$$

其中:符号 T 表示矩阵或向量的转置,列向量 \mathbf{w} 转置后成为行向量 \mathbf{w}^T ,有 $\mathbf{w}^T \mathbf{w} \in R$ 。C 是惩罚因子,也可看作 SVM 的模型复杂度 and 经验风险的平衡因素。 $\Phi(\cdot)$ 是一个映射函数,将原始空间的向量映射到一个高维(可能无限)特征空间,以解决线性不可分问题。松弛变量 $\xi \in R, k=1$ (一次松弛)、2 (二次松弛)是最常见的两种 SVM 误差度量形式。权向量 \mathbf{w} 和偏置量 b 共同构成分类函数 $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T \Phi(\mathbf{x}) + b)$ 。显然,当

$c_{+1}(\mathbf{x}_i) = c_{-1}(\mathbf{x}_i) = 1$ 时, 式(1)、(2) 退化为标准的 SVM, 两者为不同的常数时, 式(1)、(2) 为类别相关的代价敏感 SVM^[2-3]。而我们研究的目标是当 $c_{+1}(\mathbf{x}_i)$ 和 $c_{-1}(\mathbf{x}_i)$ 随样本而变化时, 生成的分类器使得测试集上误分类代价 = $\sum_{\mathbf{x}_i \in \text{测试数据集}} c_{y_i}(\mathbf{x}_i)$ 最小。为简化问题, 本文约定: 将实际类别为正例 s (负例 s) 的样本判为正例 (负例) 时, 损失为 0, 而误判为负例 (正例) 时, 损失为 $c_{+1}(s)(c_{-1}(s))$, 实际应用中, $c_{+1}(s)$ 和 $c_{-1}(s)$ 都是大于零的实数。

1.1 代价样本相关的一次松弛 csSVM

当 $k = 1$ 时, 通过构造 Lagrange 方程, 得到优化问题(1) (2) 的对偶形式^[5]:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)) \quad (3)$$

约束条件为:

$$\begin{cases} \sum_{i=1}^m \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C c_{+1}(\mathbf{x}_i), \quad \forall i \in I_{+1} \\ 0 \leq \alpha_i \leq C c_{-1}(\mathbf{x}_i), \quad \forall i \in I_{-1} \end{cases} \quad (4)$$

如何解上述对偶问题, 文献[5] 并没有介绍, 而 SVM^{Light} 提供了完整的解法^[6]。一般来说, 无法确定映射函数 $\Phi(\cdot)$, 选择核函数 $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ 反映向量 $\mathbf{x}_i, \mathbf{x}_j$ 在映射后的高维空间上的内积。

1.2 样本相关的二次松弛代价敏感 SVM

当 $k = 2$ 时, 如何实现样本相关的代价敏感 SVM, 就我们所知, 还未见报道。若有矩阵 $\mathbf{A} \in R^{m \times n}$ 、 $\mathbf{B} \in R^{n \times d}$, 记 \mathbf{A}_i 是矩阵 \mathbf{A} 的第 i 行元素形成的行向量, 依照通用 SVM (Generalized SVM, GSVM) 的通用核定义^[7], 核 $K(\mathbf{A}, \mathbf{B})$ 映射 $R^{m \times n} \times R^{n \times d}$ 为 $R^{m \times d}$, 核矩阵 \mathbf{K} 无须对称, 其元素 $K_{ij} = K(\mathbf{A}_i, \mathbf{B}_j)$ 。基于 GSVM 框架, 式(1)、(2) 可转换为以下优化问题:

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} (\mathbf{w}^T \mathbf{w} + b^2) + C \sum_{i \in I_{+1}} c_{+1}(\mathbf{x}_i) \xi_i^2 + C \sum_{i \in I_{-1}} c_{-1}(\mathbf{x}_i) \xi_i^2 \quad (5)$$

约束条件为:

$$\begin{cases} y_i (\sum_{j=1}^m K(\mathbf{x}_i, \mathbf{x}_j) y_j w_j + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \\ i = 1, 2, \dots, m \end{cases} \quad (6)$$

w_i 是权向量 \mathbf{w} 的第 i 项。在式(5) 中, 为方便后文的算法, 加入项 $0.5 \times b^2$, 这等于训练数据增加了一个为常数的属性。实际上, 当 $c_{+1}(\mathbf{x}_i) = c_{-1}(\mathbf{x}_i) = 1$ 时, 有该项对分类器的准确率影响很小^[7]。依此求得的分类超平面为:

$$\sum_{i=1}^m K(\mathbf{x}, \mathbf{x}_i) y_i w_i + b = 0 \quad (7)$$

利用正号函数 $(x)_+ = \max(0, x)$, 式(5)、(6) 可合并为无约束问题:

$$\min_{\mathbf{w}, b} \frac{1}{2} (\mathbf{w}^T \mathbf{w} + b^2) + C \sum_{i=1}^m c_{y_i}(\mathbf{x}_i) (1 - y_i (\sum_{j=1}^m K(\mathbf{x}_i, \mathbf{x}_j) y_j w_j + b))^2_+ \quad (8)$$

因为 $c_{+1}(\mathbf{x}_i) > 0, c_{-1}(\mathbf{x}_i) > 0$, 其余项皆为二次项, 显然, 式(8) 是严格凸的无约束最优化函数, 有唯一的最优解。但是正号函数不可微, 为能使用快速求解的 Newton 算法, 降低 SVM 的复杂性, 引入分段多项式平滑函数 $p(x, \delta)$ 近似正号函数^[9]:

$$p(x, \delta) = \begin{cases} x, & x \geq \delta \\ -\frac{1}{16\delta^3} (x + \delta)^3 (x - 3\delta), & -\delta < x < \delta \\ 0, & x \leq -\delta \end{cases} \quad (9)$$

$p(x, \delta)$ 具有二阶光滑性, 在区间 $(-\delta, +\delta)$ 能很好地逼近 $(x)_+$, 两者间的误差^[9]: $p(x, \delta)^2 - (x)_+^2 \leq \frac{\delta^2}{19} \approx 0.0526\delta^2$ 。

当样本个数 m 很大时, 核计算非常耗时, 为此, 简约核 (reduced kernel)^[10-11] 在 SVM 得到应用。按照一定的规则从训练集采集 $d \ll m$ 个样本, $\{(\mathbf{x}'_i, y'_i) \mid i = 1, \dots, d\}$ 。用简约核 K' 代替满核 K , 核矩阵从 $m \times m$ 减少到 $m \times d$, 式(8) 变为:

$$\min_{\mathbf{w}, b} \frac{1}{2} (\mathbf{w}^T \mathbf{w} + b^2) + C \sum_{i=1}^m c_{y_i}(\mathbf{x}_i) (1 - y_i (\sum_{j=1}^d K(\mathbf{x}_i, \mathbf{x}'_j) y'_j w_j + b))^2_+ \quad (10)$$

相应的分类超平面为:

$$\sum_{i=1}^d K(\mathbf{x}, \mathbf{x}'_i) y'_i w_i + b = 0 \quad (11)$$

式(8) 和(9) 合并后, 获得优化问题:

$$\min_{\mathbf{w}, b} \psi_{\delta}(\mathbf{w}, b) := \min_{\mathbf{w}, b} \frac{1}{2} (\mathbf{w}^T \mathbf{w} + b^2) + C \sum_{i=1}^m c_{y_i}(\mathbf{x}_i) (p(1 - y_i (\sum_{j=1}^d K(\mathbf{x}_i, \mathbf{x}'_j) y'_j w_j + b), \delta))^2 \quad (12)$$

式(12) 和(7) 组成非简约的代价样本相关 csSVM, 记为 sd2sSVM-a, 式(10) 和(9) 合并后, 获得优化问题:

$$\min_{\mathbf{w}, b} \psi_{\delta}(\mathbf{w}, b) := \min_{\mathbf{w}, b} \frac{1}{2} (\mathbf{w}^T \mathbf{w} + b^2) + C \sum_{i=1}^m c_{y_i}(\mathbf{x}_i) (p(1 - y_i (\sum_{j=1}^d K(\mathbf{x}_i, \mathbf{x}'_j) y'_j w_j + b), \delta))^2 \quad (13)$$

式(13) 和(11) 组成简约的代价样本相关 csSVM, 记为 sd2sSVM-b, 两者统称为 sd2sSVM。

2 sd2sSVM 的 Newton-YUAN 算法

由于式(12) 和(13) 的优化函数 $\psi_{\delta}(\mathbf{w}, b)$ 具有二次可微性, 可结合 Newton 法和 YUAN^[12] 的一维精确搜索算法求解 sd2sSVM, 我们称之为 Newton-YUAN 算法。具体步骤如下:

1) 初始迭代点为 $(w^{(1)}, b^{(1)})$, 记 $k = 1$ 。多项式平滑参数 $\delta = \sqrt{\frac{2\varepsilon_2}{0.0526m}}$ ^[9], 其中 m 是训练集样本个数, ε_2 是式(12) 或(13) 的近似解与式(8) 或(10) 最优解的最大误差。

2) 计算式(12) 或(13) 优化函数的梯度 $\mathbf{g}_k = \nabla \psi_{\delta}(\mathbf{w}^{(k)}, b^{(k)})$ 。记 $\|\cdot\|_v$ 为变量的 v 范式, 如 $\|(\mathbf{w}, b)\|_2 = \sqrt{(\mathbf{w}, \mathbf{w}) + b^2}$, 如果 $\|\mathbf{g}_k\|_2 \leq \varepsilon_1$, 则停止, $(\mathbf{w}^{(k)}, b^{(k)})$ 为分类超平面式(7) 或(11) 的参数; 否则继续后面的步骤。

3) 由 Hesse 矩阵和梯度计算迭代方向 $d^{(k)}$, 并用 Cholesky 分解法解下面的线性方程组:

$$\nabla^2 \psi_{\delta}(\mathbf{w}^{(k)}, b^{(k)}) d^{(k)} = -\nabla \psi_{\delta}(\mathbf{w}^{(k)}, b^{(k)}) \quad (14)$$

4) YUAN 的最速下降法步长选择法^[12] 计算 Newton 法迭代步长, 若 $\text{mod}(k, 3) = 0$, 则:

$$(\mathbf{w}^{(k+1)}, b^{(k+1)}) = (\mathbf{w}^{(k)}, b^{(k)}) + \frac{d^{(k)} \times 2}{\sqrt{4 \times \|\mathbf{g}_k\|_2^2 / \|(w^{(k+1)} - w^{(k)}, b^{(k+1)} - b^{(k)})\|_2^2 + 2}} \quad (15)$$

式(15) 没有化简, 方便与文献[12] 的式(4.1) 对照; 反之, $\text{mod}(k, 3) \neq 0$, 则 $(\mathbf{w}^{(k+1)}, b^{(k+1)}) = (\mathbf{w}^{(k)}, b^{(k)}) + d^{(k)}$ 。令 $k = k + 1$, 继续 2)。

3 实验结果及分析

第 1, 2 章从理论上分析了 sd2sSVM 的可行性, 可用来实现二次松弛的代价样本相关 csSVM。现在, 我们在多个公开数据集上实验, 与一次松弛的代价样本相关 csSVM 比较。表

1 是实验数据集的介绍,所有数据集都是从 <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets> 下载而来,属性全部归一化到 $[-1, 1]$ 。其中带 * 号的数据集直接下载,未作任何加工,而未带 * 号者,经过了归一化(若值已经处于 $[-1, 1]$ 间的属性,则不做处理)、样本数调整(考虑正例与负例的平衡)的处理。按照 $\cos t_i = \frac{2}{1 + e^{y_i(x_{i1} + x_{i2})}}$ 赋以每个样本的代价, x_{i1} 和 x_{i2} 分别是样本 i 的第 1 维和第 2 维属性值。

表 1 实验数据集

名称	来源	样本数	属性维数	正例数
australian *	statlog	690	14	307
diabetes *	UCI	768	8	500
german *	statlog	1000	24	300
heart *	statlog	270	13	120
ionosphere *	UCI	351	34	225
a7a	UCI	16 100	122	3 918
a7a-B	UCI	9 104	122	3 911
splice. t	Delve	2 175	60	1 131
ijcnn1-B	DP01a	8 193	22	796
ijcnn1-D	DP01a	6 146	22	2 049

SVM^{Light}是我们能找到的实现一次松弛代价样本相关 csSVM 的唯一公开工具,除非特别声明,都以 SVM^{Light}表示该分类器,后文不再赘述。文章所有实验在 Pentium 2.66 GHz CPU、512 MB 内存的机器上完成,算法用 Visual C++ 6.0 实现。所有实验全部以十折交叉完成,按照数据集正例和负例的原始比例,随机地将样本平均分配成十份,将九份作为训练数据,另一份作为测试数据,取十次测试结果的平均值作为实验结果。Newton-YUAN 算法的参数 $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-3}$, SVM^{Light}的参数全部采用缺省值。所有表的度量指标单位统

一,训练精度和测试精度是百分比,训练时间以秒为单位,而训练代价和测试代价无单位。其中训练代价为 $\sum_{x_i \in \text{训练数据集}}$

$\cos t_i$,测试代价为 $\sum_{x_i \in \text{测试数据集}} \cos t_i$ 。

3.1 线性可分问题的分类比较

对于线性可分的代价样本相关问题,使用 sd2sSVM-a 分类,式(12)变为:

$$\min_{w, b} \psi_\delta(w, b) := \min_{w, b} \frac{1}{2} (w^T w + b^2) + C \sum_{i=1}^m c_{y_i}(x_i) (p(1 - \gamma_i(x_i^T w + b), \delta))^2 \quad (16)$$

分类超平面为 $x^T w + b = 0$ 。一次松弛代价样本相关 csSVM 的式(3)变为:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i, j=1}^m \alpha_i \alpha_j y_i y_j (x_i^T x_j) \quad (17)$$

分类超平面同样为 $x^T w + b = 0$, w 和 b 可由式(17)的最优解 α^* 求得。

表 2 是两种分类器针对线性可分问题的比较,惩罚因子 $C = 1$ 。由于 sd2sSVM-a 采用 Newton-YUAN 算法直接求解,4~6 次迭代即可找到最优值,比 SVM^{Light}的训练时间快 10 倍以上。当 δ 足够小时,分段多项式平滑函数 $p(x, \delta)$ 能很好地近似正号函数。而 SVM^{Light}使用解组策略和 shinking 技术避免在内存保留过多数据,优化过程数据子集的选择可能导致信息损失,从而测试代价略高于 sd2sSVM。当调整 SVM^{Light}的参数后,两者的测试代价靠近,但这种参数调整工作比较麻烦。注意,训练集大小是测试集的 9 倍,所以训练代价也差不多是测试代价的 9 倍。代价样本相关 csSVM 的优化目标是使测试代价减小,训练精度和测试精度与样本代价函数有关,不能作为分类器性能的比较指标,在表中列出,仅为参考。

表 2 线性可分问题比较

数据集	sd2sSVM-a					SVM ^{Light}				
	训练精度	训练代价	训练时间	测试精度	测试代价	训练精度	训练代价	训练时间	测试精度	测试代价
A7a	75.44	1 314.2	10.19	75.17	150.2	75.18	1 332.0	270.60	74.73	154.0
Ijcnn1_B	97.24	184.4	0.24	97.16	21.1	96.96	195.6	2.55	96.90	22.1
Ijcnn1_D	97.36	129.7	0.19	97.30	14.8	96.74	154.7	2.08	96.68	17.5

3.2 高斯核 SVM 的分类比较

实际应用中的大多数分类问题都是线性不可分的,核函数是解决问题的有效方法。实验过程中,采用最常见的高斯核 $K(x, z) = e^{-\gamma \|x - z\|_2^2}$, 参数 $\gamma = 0.1$ 。我们只关注分类器的比较,由于惩罚因子 C 和高斯核参数 γ 共同影响算法的性能,因此不考虑参数对分类器的影响, sd2sSVM-a 和 SVM^{Light}都是采用满核的算法, C 统一为 1。

sd2sSVM-b 是采用简约核的算法,由于简约核的使用,导致核矩阵部分信息丢失,松弛变量减小,故而取较大的惩罚因子, $C = (m/d)^{0.5}$, d 是简约集样本个数, m 是所有训练样本个数。sd2sSVM-b 还有一个参数 d , 相当于 SVM 的支持向量个数 SVs, 简约集和支持向量都需要保留,用来和待分类对象(测试样本)产生内积,辅助分类, d 或 SVs 越小,待分类对象的类别预测时间就越短。根据数据集样本个数, d 取 m 的 1/10、1/20 或 1/40, 保证简约集不至于过小。简约集 d 个样本的采样规则为:

1) 令 $ratio = m/d$ 是简约集比例, m_{+1} 和 m_{-1} 分别是正例和负例的样本个数, 则简约集正例样本数 $d_{+1} = m_{+}/ratio$, 负例样本数 $d_{-1} = m_{-}/ratio$;

2) 按照样本索引顺序,在 $I_{+1}(I_{-1})$ 中每 $ratio$ 个样本中随机抽取一个作为简约集的正例(负例)样本。

表 3~4 是高斯核的分类情况比较, N/A 说明 sd2sSVM-a 时间复杂性太高,无法计算得出结果。sd2sSVM-b 与 sd2sSVM-a 相比,尽管训练代价要高,但误分代价却不一定大,说明简约核虽然丢弃了一些训练信息,但也部分抵消了过学习带来的影响;而训练时间 sd2sSVM-b 则快很多,样本个数 m 很大时, sd2sSVM-a 几乎无法训练成功。总的来说,在代价样本相关 csSVM 的实际应用过程,简约核可完全代替满核。

数据集带“+”,表明在该数据集上的 sd2sSVM-b 测试代价要小于 SVM^{Light},带“=”表示两者相当,否则不如 SVM^{Light}。sd2sSVM-b 与 SVM^{Light}相比,除 ijcnn1-B 和 ijcnn1-D 外,在其他数据集上训练时间后者一般在前者的两倍以上,而这两个数据集可能核矩阵特征值的极大极小值比较靠近,导致 Newton-YUAN 的收敛速度过慢,不过暂时还难以作出理论分析。另外, SVM^{Light}恰好在这两个数据集有较少的支持向量,收敛速度相对快一些。与 sd2sSVM-b 相比,所有数据集上 SVM^{Light}的支持向量都很多,大大延长了未知类别样本的预测时间,这在一些实时分类任务中是不现实的,而 sd2sSVM-b 的 d 显著小于 SVs, 预测时间显著快于 SVM^{Light}。比较测试代价、测试精度(为节约篇幅,在表 3 中省略),两个算法则各有千秋,说明不同的数据集适合不同阶的松弛变量,这和非代价敏感 SVM

的研究结果相同,至今为止,还没有文献报道一次松弛的标准 SVM 和二次松弛的标准 SVM 之间谁的分类精度更高。

表 3 三类 SVM 高斯核的性能比较

数据集	sd2sSVM-b			sd2sSVM-a			SVM ^{Light}		
	训练代价	训练时间	测试代价	训练代价	训练时间	测试代价	训练代价	训练时间	测试代价
heart ⁺	29.57	0.01	4.04	22.10	0.62	4.20	24.23	0.05	4.79
ionosphere	15.78	0.03	2.62	6.18	1.46	1.69	8.94	0.08	1.78
australian ⁺	61.17	0.13	7.97	55.68	10.97	7.91	67.41	0.30	8.39
diabetes	130.20	0.15	15.21	126.40	13.07	14.98	127.90	0.50	14.85
german ⁺	141.70	0.38	18.20	94.62	35.16	19.68	110.00	0.52	18.60
splice. t ⁻	41.53	1.57	4.73	40.45	342.10	4.74	42.60	4.18	4.73
ijcnn1-B ⁺	518.60	34.05	57.91	N/A	N/A	N/A	652.80	10.18	72.53
ijcnn1-D ⁺	137.30	19.30	16.22	N/A	N/A	N/A	300.00	11.04	33.94
a7a_B	817.70	68.18	98.64	N/A	N/A	N/A	381.80	161.70	98.52

表 4 支持向量个数的比较

数据集	sd2sSVM-b (d)	SVM ^{Light} (SVs)
heart	27	132.9
ionosphere	35	129.8
australian	69	252.4
diabetes	76	379.7
german	100	588.4
splice. t	108	1899.6
ijcnn1-B	204	1359.1
ijcnn1-D	153	1899.8
a7a_B	227	5654.0

4 参数 C 影响简约核

我们选择了两个数据集来评估惩罚因子 C 对简约核 sd2sSVM-b 的影响,较少样本的 australian 和较多样本的 a7a-B,简约集比例 $ratio$ 分别为 20 和 40,高斯核参数 $\gamma = 0.1$ 。表 5 是 C 对 australian 的影响,表 6 是 C 对 a7a-B 的影响。

随着 C 增大,两个数据集的训练时间却拉长。这与 sd2sSVM-b 的算法有关,注意到式(13), C 增大时,函数 $\min_{w,b} \psi_\delta(w,b)$ 的值也增大, $\nabla \psi_\delta \leq \varepsilon_1$ (Newton-YUAN 的迭代终止条件)收敛变慢,自然迭代次数增多,导致训练时间拉长。可采用如下解决方法,如 $C > 1$,则优化目标除以 C :

$$\min_{w,b} \psi_\delta(w,b) := \min_{w,b} \frac{1}{2C} (w^T w + b^2) + \sum_{i=1}^m c_{y_i}(x_i) (p(1 - y_i(\sum_{j=1}^d K(x_i, x'_j) y'_j w_j + b), \delta))^2 \quad (18)$$

由于 C 是常数,式(13)的分类精度和代价并不会改变;如 $C \leq 1$,则仍用式(13)。

表 5 C 对 australian 的影响 ($ratio = 20$)

C	训练精度	训练代价	训练时间	测试精度	测试代价
0.1	86.83	70.18	0.095	86.40	8.24
0.3	86.89	68.14	0.106	86.10	8.08
0.9	86.83	66.35	0.116	86.53	7.77
2.7	87.34	61.75	0.122	86.24	7.65
8.1	87.47	60.17	0.130	86.10	7.76
24.3	87.76	58.08	0.130	86.53	7.57
72.9	88.05	55.77	0.138	86.38	7.38
218.7	88.50	52.68	0.142	85.80	7.73

两个数据集上,较大的 C 往往使得测试代价减小,验证了前面的猜测。简约核代替满核,促使松弛变量 ξ 减小,故而要增大惩罚因子 C 来补偿 ξ ,然而简约集样本是随机产生的, ξ 要

补偿多少,无法确定,何况分类器的性能还与高斯核参数 γ 有关。与标准 SVM 一样,过大的 C 又导致过学习,这从另一个侧面可以得到映证,训练代价随着 C 增大而单调减小,因此增大 C 并不能总使测试代价降低。a7a-B 和 australian 略有不同,前者的测试代价随 C 增大而单调减小,该数据集的最优 C 是一个很大的数值。而对 australian, C 的选择更复杂,需要更多的实验才能找到最优 C ,显然参数 C 与测试代价并非线性关系。如何选择参数 C ,值得进一步研究。

表 6 C 对 a7a-B 的影响 ($ratio = 40$)

C	训练精度	训练代价	训练时间	测试精度	测试代价
0.1	69.14	1062.2	46.88	68.37	123.50
0.3	70.99	969.4	54.07	70.29	112.50
0.9	72.80	890.1	60.44	71.90	105.40
2.7	73.68	845.2	64.53	72.58	102.00
8.1	74.70	812.2	70.15	73.50	99.90
24.3	75.40	798.2	77.27	74.02	99.03
72.9	75.43	791.4	82.95	74.34	98.10
218.7	75.87	773.5	83.55	74.65	96.90

5 结语

标准 SVM 假设所有样本代价相同,代价敏感 SVM 假设所有类别的误分类代价相同,而现实世界有一些分类问题,该假设并不成立,错误分类不同样本导致的损失并不相同,因此,代价与样本相关的 csSVM 成为新的研究课题。一次松弛的代价样本相关 csSVM,虽然能够减小误分类代价,但支持向量数目过大,无法适应实时性高的分类任务。我们在平滑 SVM^[7] 和简约核^[8] 的基础上提出 sd2sSVM,支持二次松弛的代价样本相关 csSVM,同样能减小误分类代价,且分类精度也相当,而训练时间和预测时间更短,更适合那些实时分类系统。但简约核的引入,又产生了新的参数简约集大小 d ,如何使得惩罚因子 C 和 d 匹配,将是未来要解决的问题。

参考文献:

- [1] 沈曙光,王广军,朱丽娜. 基于支持向量机的逆动力学模型辨识及应用[J]. 系统仿真学报, 2008, 20(1): 25-28.
- [2] 郑恩辉,李平,宋执环. 代价敏感支持向量机[J]. 控制与决策, 2006, 21(4): 473-476.
- [3] CAMPARELLI P, CASIRAGHI E, VALENTINI G. Support vector machines for candidate nodule classification[J]. Neurocomputing, 2005, 68(10): 281-288.
- [4] ZADROZNY B, ELKAN C. Learning and making decisions when costs and probabilities are both unknown[C]// Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001: 204-213.

(下转第 2880 页)

从表 1 我们可以发现,本算法能够快速、准确地寻找到最优路径,虽然在某些算例中平均消耗时间不如 ACO-grid 和 RRT 算法短,这可能是由于对可选方向上最大信息素的搜索造成的,但是可行解的平均长度却是所有算法中最稳定,最精

确的。

为了不失一般性,我们又随机测试了一些规模较大的算例,每个算例测试 10 次,性能如表 2 所示,图 4 为算法对规模为 50×50 栅格测试得到的结果。

表 1 算法性能对比

问题规模	问题最优解长度	可行解平均长度					可行解比例/%					平均消耗时间				
		A *	GA	ACO-grid	RRT	本文算法	A *	GA	ACO-grid	RRT	本文算法	A *	GA	ACO-grid	RRT	本文算法
10 × 10	19	19	19.0	19.0	N/A	19	100	100	100	N/A	100	1.6	1.9	< 0.1	N/A	0.46
25 × 30	39	N/A	50.0	39.3	N/A	39	0	35	100	N/A	100	>300	38.0	12.6	N/A	1.44
20 × 20	54	N/A	69.8	58.7	N/A	54	0	65	95	N/A	100	>300	62.5	24.4	N/A	2.20
30 × 30	59	N/A	85.9	65.0	61.2	59	0	10	90	100	100	>300	107.1	41.7	2.992	4.39

表 2 较大规模算例测试

算例	最优解	障碍栅格覆盖率/%	可行解		平均消耗时间
			平均长度	比例/%	
40 × 30	68	20	68.0	100	4.41
30 × 50	78	23	79.1	100	8.05
40 × 40	78	25	78.0	100	7.51
50 × 50	98	23	99.7	100	10.75

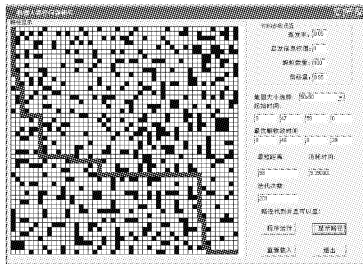


图 4 50×50 栅格图测试例子

由以上实验可以看出,与其他算法相比,本算法显示出了良好的快速求解性能,并且在规模较大、环境复杂的情况下,解的质量也显示了较高的稳定性。此外,通过试验,我们发现起始点和目标点之间只要有一条通道客观存在,不管路径多么复杂,本算法都能迅速地规划出优化路径。

4 结语

一般蚁群算法在障碍物复杂、起始点与目标点相距比较远的情况下,如果只是依靠信息素和距离信息是很难找到一条从起始点到目标点的有效路径的,并且容易陷入“死锁”。本文提出一种基于栅格模型的改进蚁群算法,它采用一种折返的迭代方式,使蚂蚁从正反两个方向对最优路径进行搜索,保证了搜索的多样性,使搜索不易陷于停滞。算法还针对问

题的特点,重构了信息素的更新方式和撒播方式,使信息素对最优路径的反应更为迅速;此外,使用可选方向范围内最大信息素和目标引导函数作为启发式因子,使蚂蚁对最优路径更为敏感,更好地适应信息素环境。仿真结果表明,与其他算法相比,该算法具有速度快、稳定性高、效果好等特点,即使在复杂的环境下,也可以快速地规划出一条优化路径。

参考文献:

- [1] 国海涛,朱庆保,徐守江.基于栅格法的机器人路径规划快速搜索随机树算法[J].南京师范大学学报:工程技术版,2007,7(2): 58-61.
- [2] 张美玉,黄翰,郝志峰,等.基于蚁群算法的机器人路径规划[J].计算机工程与应用,2005,41(9):34-37.
- [3] 朱庆保,张玉兰.基于栅格法的机器人路径规划蚁群算法[J].机器人,2005,27(2):132-136.
- [4] 朱庆保,杨志军.基于变异和动态信息素更新的蚁群优化算法[J].软件学报,2004,5(2):185-192.
- [5] 秦元庆,孙德宝,李宁,等.基于粒子群算法的移动机器人路径规划[J].机器人,2004,26(3):222-225.
- [6] 樊长虹,陈卫东,席裕庚.动态未知环境下一种 Hopfield 神经网络路径规划方法[J].控制理论与应用,2004,26(3):345-350.
- [7] 刘天孚,程如意.带精英策略和视觉探测蚁群算法的机器人路径规划[J].计算机应用,2008,28(1):92-96.
- [8] CAI ZI-XING, PENG ZHI-HONG. Cooperative coevolutionary adaptive genetic algorithm in path planning of cooperative multi-mobile robot systems[J]. Journal of Intelligent and Robotic Systems, 2002, 33(1):61-71.
- [9] SURMANN, H, HUSER J, WEHKING J. Path planning for a fuzzy controlled autonomous mobile robot[C]// Proceedings of the 5th IEEE International Conference on Fuzzy Systems. New Jersey: IEEE Press, 1996:1660-1665.

(上接第 2866 页)

- [5] BREFELD U, GEIBEL P, WYSOTSKI F. Support vector machines with example dependent costs[C]// Proceedings of the 14th European Conference on Machine Learning. Berlin: Springer-Verlag, 2003:23-34.
- [6] JOACHIMS T. Making large-scale SVM learning practical[M]// SCHÖLKOPF B, BURGESS C J C, SMOLA A J. Advances in Kernel Methods-Support Vector Learning. Cambridge: MIT Press, 1999:169-184.
- [7] LEE Y J, MANGASARIAN O L. SSVM: A smooth support vector machine[J]. Computational Optimization and Applications, 2001, 20(10):5-22.
- [8] MANGASARIAN O L. Generalized support vector machines[M]//

- SMOLA A, BARTLETT P, SCHÖLKOPF B, et al. Advances in Large Margin Classifiers. Cambridge: MIT Press, 2000:135-146.
- [9] 袁玉波,严杰,徐成贤.多项式光滑的支撑向量机[J].计算机学报,2005,28(1):9-17.
- [10] LEE Y J, HUANG S Y. Reduced support vector machines: A statistical theory[J]. IEEE Transactions on Neural Networks, 2007, 18(1):1-13.
- [11] LIN K M, LIN C J. A study on reduced support vector machines[J]. IEEE Transactions on Neural Networks, 2003, 14(6):1449-1459.
- [12] YUAN YA-XIANG. Step-sizes for the gradient method[C]// Proceedings of the 3rd International Congress of Chinese Mathematicians. Hong Kong: [s. n.], 2004:785-796.