

增量支持向量机算法研究

孙名松,张立新,杜春燕

(哈尔滨理工大学 网络中心,黑龙江 哈尔滨 150080)

摘要:在进行增量学习时,随着新增样本的不断加入,致使训练集规模不断扩大,消耗大量计算资源,寻优速度缓慢。在深入研究了支持向量分布的特点的基础上提出了分治加权增量支持向量机算法。该算法有效利用了广义 KKT 条件和中心距离比值,舍弃对后续训练影响不大的样本,得到边界支持向量集,对训练样本进行有效的淘汰。将所剩样本合并,进行加权处理,解决某些样本严重偏离所属的类别,对正常分布的样本不公平的问题。实验结果表明,该方法在保证分类精度的同时,能有效地提高训练速度。

关键词:支持向量机;增量训练;中心距离比值;加权算法

中图分类号:TP391.4

文献标识码:A

文章编号:1673-629X(2011)05-0040-04

Research on Increasing Support Vector Machine Algorithms

SUN Ming-song, ZHANG Li-xin, DU Chun-yan

(Network Information Center, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: When carried on the increase studies, along with additional sample's unceasing joined, cause the training regulations scale unceasingly expanding, consumption of massive calculation resources, and the optimization speed is slow. Propose the partitioning weighting increase support vector machines algorithm in the deep research support vector distributed characteristic's foundation. This algorithm has effectively used the generalized KKT condition and center distance ratio, discards to the training sample that is not big influence, obtains the boundary support vector collection, carries on the effective elimination to the training sample. Then merge remain the sample, carries on weighting processing to solve the question that the certain sample serious deviate respective category, regarding normal distribution sample unfair question. The experimental result indicated that this method can guarantee classification precision, can raise the training speed effectively at the same time.

Key words: support vector machine; incremental training; center distance ratio algorithm; weighted algorithm

0 引言

支持向量机(Support Vector Machines)是建立在统计学习理论^[1] VC 维理论和结构风险最小原理基础上的针对小样本的机器学习方法,是数据挖掘中的一项新技术,它能够根据有限的样本信息,借助于最优化方法解决机器学习问题,在处理高维数据时,有效地解决了“维数灾难”问题。与人工智能现有的机器学习方法相比,支持向量机具有较好的非线性处理能力和推广能力。在模式识别和回归估计等问题中已得到广泛应用^[2]。

但是对于传统的支持向量机训练算法,如果有新增样本加入,就需要对所有训练样本重新进行训练,这就要消耗大量的运算资源,给模型的推广能力带来巨

大的限制。如果既能减少新样本的加入重新学习的时间,又能继承之前所学习的知识,那么就应该将原有的分类信息保存起来与新加入的样本一起进行训练。文献[3]提出一种使用多支持向量机进行增量学习的算法,但由于初始训练是在全部的初始训练集上进行,训练速度的提高并不明显。该方法只解决了在大样本情况下的 SVM 学习问题。文献[4]提出了一种 v-svm 支持向量机增量学习算法,其训练集的获得主要从支持向量、误分数据中有选择地淘汰一些样本,同时该算法有多个参数需要选择,但却没有确定这些参数的一个行之有效的方法。文献[5]中的 CDRM+SVM 方法是先利用中心距离比值方法抽取出支持向量样本,然后再对训练样本进行优化。但是该方法在实际操作时涉及到两个阈值的选取问题,而且与抽取的向量息息相关,如果阈值选取的过大,则无法包括全部支持向量;若阈值选取过小,则抽取的边界向量过多。因此由于阈值的选取缺乏一个良好的标准,导致该方法实用性不强。文献[6]采用加权的思想来控制“野点”对分类

收稿日期:2010-10-07;修回日期:2011-01-13

基金项目:黑龙江省自然科学基金(F9608)

作者简介:孙名松(1963-),男,教授,研究方向为网络安全、网络应用;张立新,硕士研究生,研究方向为网络安全、模式识别。

超平面的影响,考虑了由于样本分布不同带来的权值差异,保留尽可能多的样本类别信息。但在数据集训练中,如果得到的支持向量比较多,那么该算法在筛选样本、缩减数据规模的时候不是很有效,造成训练时间不理想。

文中受到上述增量学习算法、中心距离比值法及加权算法的启发,结合支持向量只占训练样本比较少的比例的特点,分析了中心距离比值和加权值对支持向量集的影响,提出了分治加权增量支持向量机算法。

1 边界向量条件分析

由于最优分类面只由支持向量确定,所以只需求出支持向量便可求出最优分类面。并不是所有的训练样本都可能成为支持向量。因此为了提高算法的速度,可首先对训练样本进行选择,选择那些有可能成为支持向量的边界样本,并用这些边界向量训练。这样便可在保证算法精度的同时,减少参与训练的样本数,从而提高算法的速度。

1.1 KKT 条件分析

支持向量机对应的 KKT 条件为^[7,8]:

$$\begin{cases} \alpha_i = 0 \Rightarrow f(x_i) \geq 1 \text{ 或 } f(x_i) \leq -1 \\ 0 < \alpha_i < c \Rightarrow f(x_i) = 1 \text{ 或 } f(x_i) = -1 \\ \alpha_i = c \Rightarrow -1 \leq f(x_i) \leq 1 \end{cases} \quad (1)$$

通过(1)式可以看出如果对应的样本在分类器分类间隔之外,那么 $\alpha_i = 0$;如果对应的样本位于分类间隔之上,那么 $0 < \alpha_i < C$;其他则 $\alpha_i = C$ 。那么满足 KKT 条件为:

$$y f(x_i) \geq 1 \quad (2)$$

即:满足 KKT 条件的样本,为那些位于分类间隔之外且被分类器正确分类的样本和支持向量。

违背 KKT 条件的样本则可以分为以下三类:

1) 满足 $0 \leq y f(x_i) < 1$,能够被原分类器正确分类的样本,与本类在分类边界同侧,位于分类间隔中;

2) 满足 $-1 \leq y f(x_i) \leq 0$,被原分类器错误分类的样本,与本类在分类边界异侧,位于分类间隔中;

3) 满足 $y f(x_i) < -1$,被原分类器错误分类的样本,与本类在分类间隔异侧,位于分类间隔外。

但如果新增样本和原有训练样本有样本发生重复,且新增样本刚好部分包括原有 SVM 的支持向量,虽然新增的样本满足 KKT 条件,且不会对增量学习后的最优分类面和支持向量机产生影响,但是该新增样本却成为了新的支持向量。因此,为了使算法有更宽的推广能力,可以将满足 KKT 条件放宽为:

$$y f(x_i) > 1 \quad (3)$$

称之为广义 KKT 条件。那么违背广义 KKT 条件则

为:

$$y f(x_i) \leq 1 \quad (4)$$

即是违背 KKT 条件的样本和支持向量机(SV)集的并集。

上述讨论可知,利用广义 KKT 条件对样本集进行筛选,可以舍弃那些对分类无贡献的样本,保留样本集中样本的特征信息。违背广义 KKT 条件的样本将会影响增量学习后的 SV 集。

1.2 支持向量变化情况

以图 1 为例来分析新增样本训练后 SV 集可能发生的变化^[9]。

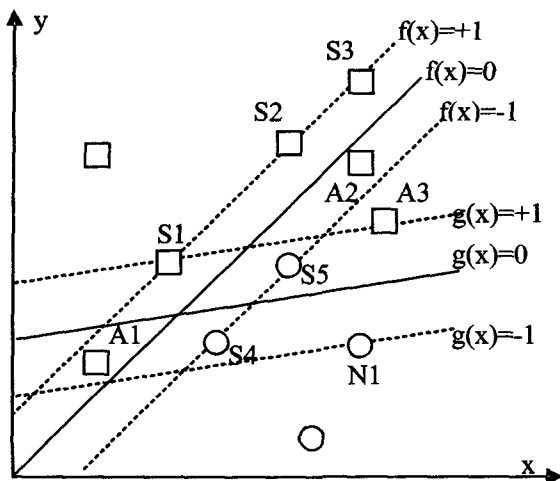


图 1 新增样本训练后 SV 集可能发生的变化

图 1 中 S1、S2、S3、S4、S5 为原始支持向量, A1、A2、A3 是新增加的样本,原始样本集的最优分类超平面为 $f(x) = 0$,对新增的样本进行训练后得到的最优分类超平面为 $g(x) = 0$ 。N1 为原样本集中的非 SV,训练后的 SV 集由 S1、A3、N1 组成。从图 1 中可以看出,新增样本 A2 和 A3 属于被错分的样本,违背原始样本集上 KKT 条件,在重新训练后转化为满足 KKT 条件的样本;且样本 A3 在训练后转化成支持向量;而原来 SV 集中的 S1、S2、S3、S4、S5 在训练后只有 S1 仍是支持向量,其他几个都转化为满足 KKT 条件的样本;原样本 N1 满足 KKT 条件在训练后转化成为 SV。从上图还可以看出,那些在原样本中的非 SV 经过训练后转化为新的 SV,或者由原来满足 KKT 条件的样本训练为新的 SV 的这两种样本,一般都分布在分类间隔附近。基于上述分析,满足广义 KKT 条件的样本中与原分类面间隔距离较近的样本,和违背广义 KKT 条件的样本在新一轮训练之后有可能转化为新的 SV。

2 分治加权淘汰算法

2.1 淘汰算法

由 1.2 节的分析知道除了要保留违背广义 KKT 条件的样本,对于包括新增样本集和原样本集中的满

足广义 KKT 条件的样本,还需要选择距离分类面间隔较近的样本。必须采取有效的方法,淘汰掉无用样本,保留重要信息。首先给出一些涉及到的定义^[10,11]。

定义 1(中心) 给定一类样本,其平均特征称为该类样本的中心。给定一类训练样本 $\{x_1, x_2, \dots, x_n\}$, 那么其中心为:

$$m = \frac{1}{n} \sum_{i=1}^n x_i \quad (5)$$

特征空间样本的中心向量要在特征空间中求得:

$$m = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \quad (6)$$

定义 2(距离) 两个样本之间的特征差异称为样本距离。已知两个 N 维样本向量 x_1 和 x_2 , 其样本距离为:

$$d(x_1, x_2) = \|x_1 - x_2\|_2 = \sqrt{\sum_{i=1}^N (x_1^i - x_2^i)^2} \quad (7)$$

两样本 x_1 和 x_2 在特征空间之间的样本距离可表示为:

$$d(\varphi(x_1), \varphi(x_2)) = \sqrt{k(x_1, x_1) - 2k(x_1, x_2) + k(x_2, x_2)} \quad (8)$$

定义 3(中心距离) 各样本到类中心的距离。且对每个正类样本点来说,都有两个中心距离:自中心距离和互中心距离。

给定训练样本

$$X^+ = \{x_i \mid x_i \in R^m, y_i = 1, i = 1, 2, \dots, n^+\}$$

$$X^- = \{x_i \mid x_i \in R^m, y_i = -1, i = 1, 2, \dots, n^-\}$$

则其中心分别为 $m^+ = \frac{1}{n^+} \sum_{i=1}^{n^+} \varphi(x_i)$, $m^- = \frac{1}{n^-} \sum_{i=1}^{n^-} \varphi(x_i)$, 对每个正类样本点来说,都有两个中心距离:自中心距离 $D_{ii} = d(\varphi(x_i), m^+)$ 和互中心距离 $D_{mi} = d(\varphi(x_i), m^-)$; 同理对每个负类样本也都有两个中心距离:自中心距离 $D_{mi} = d(\varphi(x_i), m^-)$ 和互中心距离 $D_{ii} = d(\varphi(x_i), m^+)$ 。

中心距离为:

$$d(x_i, m^+) = \|x_i - m^+\|_2 = \sqrt{\sum_{j=1}^{n^+} (x_i^j - m^j)^2} \quad (9)$$

在特征空间的中心距离为:

$$d(\varphi(x), m^+) = \sqrt{k(x, x) - \frac{2}{n^+} \sum_{i=1}^{n^+} k(x, x_i) + \frac{1}{n^+{}^2} \sum_{i=1}^{n^+} \sum_{j=1}^{n^+} k(x_i, x_j)} \quad (10)$$

定义 4(中心距离比值) 已知两类样本,求出某一个样本 x_i 的互中心距离和自中心距离,两者的比值 R 称为中心距离比值:

$$R = \frac{D_{mi}}{D_{ii}} \quad (11)$$

根据第一节分析那些有可能成为能够为边界向量的样本点是距离分类间隔较近的样本点,对应中心距离比值较小,仅占有所有训练样本的一小部分,并且其中包含支持向量样本。

采取淘汰的方法如下:首先利用 KKT 条件,对样本进行分类选择,对于违背 KKT 条件的样本,求出该类样本中的样本点对应的样本中心,利用公式求出自中心距离和互中心距离,接着利用中心距离比值公式求出中心距离比值。按照每类样本中心距离比值的大小对样本进行排序,取中心距离比值较小的(取前大约 20%)所对应的训练样本点作为该类的边界向量样本集合。这样即使中心距离比值得到了有效的利用,又解决了阈值选取的难点。

2.2 加权思想

同样眼前又摆出了另一个问题,当前各算法都对参加训练的所有样本平等对待。然而在很多实际应用中,由于噪声和其他多种不确定因素的存在,使得某些样本点出现间隔比较小的情况(这些点也叫离群点或者野点),严重偏离所属的类别,它在学习时就应该被适当的忽略,这对分类器来说是种损失。但是放弃这些点也带来了好处,那就是可以得到更大的几何间隔。显然必须权衡这些错分的样本对分类器影响程度,把错分样本和正常样本带来的不同损失区分开。由此引入加权的想法。

基于之前的分析除了保留了违背广义 KKT 条件的样本,同时对于包括新增样本集和原样本集中的满足广义 KKT 条件的样本,利用中心距离比值方法选择距离分类面间隔较近的样本。此时的边界向量集已经选取出来。

对每一个边界向量集中的样本都有它们自中心距离和互中心距离的数值记录,那么权值 Q 公式如下:

$$Q = \frac{D_{mi}}{D_{ii} + D_{mi}} \quad (12)$$

文献[12]以两类正态随机分布的样本为例,解释了这种给样本附加经验权值估计方法的合理性。对于两类的分类问题,当样本同时远离 2 个样本中心时,估计值逐渐趋近于 0.5,当样本靠近另外一个类别中心,远离所属类别中心时,估计值逐渐趋近于 0。权值的取值范围 $Q \in [0, 1]$ 。

在增加了样本的权值信息之后,得到的超平面改写为:

$$y_i((w \cdot x_i) + b) \geq d_i, i = 1, \dots, n \quad (13)$$

相应得到的二次规划问题为:

$$\left\{ \begin{array}{l} \min \frac{1}{2} \|w\|^2 \\ y_i(w \cdot x_i + b) \geq d_i \\ i = 1, \dots, n \end{array} \right\} \quad (14)$$

通过等价变换,将上式转化为下列凸二次规划问题:

$$\left\{ \begin{array}{l} \max W(a) = \sum_{i=1}^n a_i d_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j \langle x_i, x_j \rangle \\ \text{s. t. } \sum_{i=1}^n a_i y_i = 0, a_i \geq 0, i = 1, \dots, n \end{array} \right\} \quad (15)$$

设 a^*, b_0 是上述优化问题的解,则基于加权最优超平面的决策函数为:

$$f(x) = \text{sgn}(\sum_{i=1}^n a_i^* y_i (x_i \cdot x) + b_0) \quad (16)$$

$$b_0 = \frac{1}{2} [(w_0 \cdot x_i^*(1) - d_j^*(1)) + (w_0 \cdot x_i^*(-1) - d_j^*(-1))] \quad (17)$$

其中 $x^*(1), d^*(1)$ 表示属于第 1 类的一个支持向量和权值; $x^*(-1), d^*(-1)$ 表示属于第 2 类的一个支持向量和权值。

通过对样本进行分治选择和加权处理,明显缩小历史数据的存储,并解决了野点问题,直观易懂,实现方便。

3 分治加权增量训练算法描述

主要运算过程描述如下:

给定初始样本集合为 X_0 ,新增加的样本集合为 X_1 。

步骤一:对原始样本 X_0 中的样本通过分治加权算法进行训练,得到初始分类器 S_0 并根据判断结果将 X_0 分为 X_0^S (满足广义 KKT) 和 X_0^{NS} (违背广义 KKT);

步骤二:利用 S_0 的广义 KKT 条件,检验 X_1 中的样本点。若没有违背 S_0 的广义 KKT 条件,则算法停止, S_0 为增量学习结果;否则,根据判断结果将 X_1 分为 X_1^S (满足广义 KKT) 和 X_1^{NS} (违背广义 KKT),转向步骤三;

步骤三:对于新增的样本集 X_1 中的样本采用分治加权增量支持向量机算法进行训练,得到分类器 S_1 ;

步骤四:利用 S_1 的广义 KKT 条件,检验 X_0 中的样本点。违背 S_1 的广义 KKT 条件,则算法停止, S_1 为增量学习的结果;否则转向步骤五;

步骤五:将 X_0^S 与 X_1^S 合并得 X^S ,根据标识符,将集合分为正例样本集 X^+ 和负例样本集 X^- ,计算正负例样本点的中心距离比值。按淘汰规则进行处理,舍弃对后续训练影响不大的样本,得到剩余的正负例样本

集,合并得预边界向量集 X^B 。

步骤六:将 X^B, X_0^{NS}, X_1^{NS} 合并为 X_0 ,对 X_0 中的样本进行样本权值的估计,按照加权的支持向量机算法进行训练,得到最终的分类器 S_0 ;

步骤七: X_0 保存,作为下一次增量学习的初始样本。

4 实验评估

为了验证分治加权增量支持向量机分类算法的有效性,文中的实验主要从训练时间和训练精度两个方面进行考虑。所谓训练时间指的就是算法完成学习任务所消耗的时间;训练精度指的就是算法能对输入的数据进行正确、精确的的分类的程度。实验数据来自于 <http://archive.ics.uci.edu> 提供的二类别标准样本数据库 SPECT Heart,样本个数 267 个,属性 22。对样本数据集随机抽取训练样本数 80,每次的增量样本数为 47,增量次数为 4 次。与标准 SVM 和基于 KKT 条件的增量支持向量机进行对比试验,验证实验结果。实验环境为 MATLAB7.1。

从表 1 可以看出,分治加权增量支持向量机算法在训练时间上比标准 SVM 和基于 KKT 条件的增量 SVM 都要少,可以显著提高运算速度;而在训练的精度上比标准 SVM 要高很多,略低于基于 KKT 条件的增量 SVM。主要原因是该算法与其他两种算法相比可以显著减少历史数据的存储。

表 1 SPECT Heart 数据库上算法性能的对比

样本集	标准增量 SVM		基于 KKT 条件的增量 SVM		分治加权增量 SVM	
	训练时间 (ms)	训练精度 (%)	训练时间 (ms)	训练精度 (%)	训练时间 (ms)	训练精度 (%)
初始样本集	402	92	426	92	473	92
增量样本集 1	557	86	646	92	520	90
增量样本集 2	775	80	912	91	604	92
增量样本集 3	1008	83	1319	91	697	91
增量样本集 4	1211	86	1631	96	831	95

5 结束语

文中根据支持向量在样本空间的分布特性,在结合 KKT 条件及基于中心距离比值算法和样本加权思想的基础上提出了分治加权增量支持向量机算法,该算法可以对训练数据进行有效的淘汰,显著减少历史数据的存储。同时自动地对这些样本进行适应,把错分样本和正常样本带来的不同损失区分开。在不影响 SVM 的训练精度的前提下,大大减少了支持向量机训

(下转第 47 页)

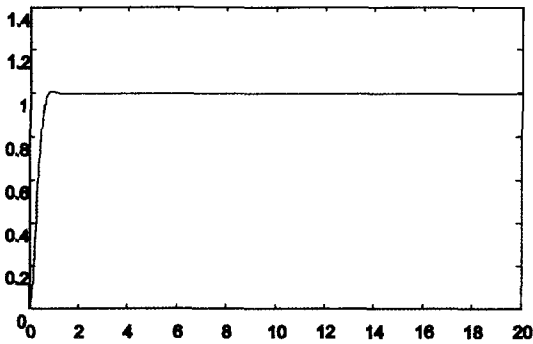


图3 基于调整函数的模糊控制系统阶跃响应曲线

4 结束语

针对某一被控对象,提出了一种基于 MATLAB 的 $\alpha(t)$ 调整函数实现及其参数优化方法,建立了基于 $\alpha(t)$ 调整函数的模糊控制系统优化的仿真分析模型。在 MATLAB 环境下搭建了仿真模型并进行了仿真。

结果表明:

(1) 基于 $\alpha(t)$ 调整函数的自适应模糊控制器的控制性能明显优于基于 α 调整因子的自适应模糊控制器。

(2) $\alpha(t)$ 调整函数的初值 α_0 及调整率 η 对 $\alpha(t)$ 调整函数的调整效果影响大。

(3) $\alpha(t)$ 调整函数的初值 α_0 及调整率 η 可以通过 MATLAB 优化工具箱的极小化函数 `fminsearch` 参数优化取得优选值。

参考文献:

[1] 佟绍成. 非线性系统的自适应模糊控制[M]. 北京: 科学出

(上接第43页)

练时间。

实验结果验证了算法的可行性和有效性。

参考文献:

- [1] Vapnik V N. Principle of Risk Minimization for Learning Theory[J]. Advances in Neural Information Processing Systems, 1992, 4(2): 831-838.
- [2] Cristianini N. An introduction to support vector machines [M]. Cambridge: Cambridge University Press, 2000: 51-54.
- [3] 萧嵘, 王继成, 孙正兴, 等. 一种 svm 增量学习算法 a-IS-VM[J]. 软件学报, 2001, 12(12): 1818-1823.
- [4] 李凯, 黄厚宽. 支持向量机增量学习算法研究[J]. 北方交通大学学报, 2003, 27(5): 34-37.
- [5] Zhang L, Zhou W D, Jiao L C. Pre-extracting support vectors for support vector machine[C]//Proceeding of ICSP2000. [s. l.]: IEEE, 2000: 1432-1435.
- [6] 王晓燕. 加权增量的支持向量机分类算法研究[D]. 杭州:

版社, 2006: 3-9.

- [2] 张茂元, 邹春燕, 卢正鼎, 等. 一种基于变调整学习规则的模糊网页分类方法研究[J]. 计算机研究与发展, 2007, 44(1): 99-1.
- [3] Tong R M, Beck M B, Latteern A. Fuzzy control of activated sludge wastewater treatment process[J]. Automat, 1980, 16(6): 695-701.
- [4] Shao S. Fuzzy self-organizing controller and its application for dynamic process[J]. Fuzzy sets and systems, 1988, 26: 151-164.
- [5] 刘曙光, 魏俊民, 竺志超. 模糊控制技术[M]. 北京: 中国纺织出版社, 2001: 135-154.
- [6] 赵纪涛, 马莉, 王现君, 等. 一种自适应的模糊关联规则挖掘算法[J]. 计算机技术与发展, 2008, 18(5): 64-67.
- [7] Deng Xiantu. Generating rules fuzzy logic controllers by functions[J]. Fuzzy sets and systems, 1990, 36: 83-89.
- [8] 冯冬青, 张希平, 费敏锐, 等. 一种基于 MATLAB 的模糊控制器综合优化设计方法[J]. 系统仿真学报, 2004, 16(4): 849-852.
- [9] 张国良, 曾静, 邓方林, 等. 模糊控制 MATLAB 应用[M]. 西安: 西安交通大学出版社, 2002.
- [10] 魏华, 李群, 陈得宝, 等. 一种新型参数非线性模糊 PID 控制方法[J]. 计算机技术与发展, 2008, 18(2): 237-243.
- [11] 曹志国, 廉小亲. 基于 MATLAB 的两种模糊控制系统的仿真方法[J]. 计算机仿真, 2004, 21(3): 41-44.
- [12] 李博, 龚晓宏. 基于 MATLAB 的模糊控制系统的优化设计与仿真[J]. 电子元器件应用, 2005, 7(3): 54-56.
- [13] Fuzzy Logic Toolbox for Use with Matlab[M]. [s. l.]: The Math Works, Inc, 1999.
- [7] 姜雪, 陶亮, 王华彬, 等. 基于分层并行筛选样本的 SVM 增量学习算法[J]. 计算机技术与发展, 2007, 17(11): 92-95.
- [8] 王晓丹, 郑春颖, 吴崇明, 等. 一种新的 SVM 对等增量学习算法[J]. 计算机应用, 2006, 26(10): 2440-2443.
- [9] 薛伟. 基于大规模训练集的 SVM 研究[D]. 秦皇岛: 燕山大学, 2009.
- [10] 孔波, 刘小茂, 张均. 基于中心距离比值增量支持向量机[J]. 计算机应用, 2006, 26(6): 1434-1437.
- [11] 徐海龙, 王晓丹, 史朝辉, 等. 一种基于距离比值的支持向量机增量训练算法[J]. 空军工程大学学报(自然科学版), 2008, 9(4): 29-33.
- [12] 鹿卫国, 戴压平, 涂序彦, 等. 适用于加权样本集处理的加权支持向量机方法[J]. 北京理工大学学报, 2005, 25(3): 211-215.