

基于数据分布规律的分段组合支持向量机研究

陈 鹏¹, 刘 爽², 左 莉³, 李锡祚²

(1 大连东软信息学院 计算机科学与技术系, 辽宁 大连 116023;

2 大连民族学院 计算机科学与工程学院, 辽宁 大连 116600;

3 大连民族学院 经济管理学院, 辽宁 大连 116600)

摘 要: 在分析数据分布规律的基础上, 提出了分段组合支持向量机算法. 该算法首先统计数据分布规律, 采用 k 均值聚类计算聚类中心, 然后分球形分布、线性二分可分、局部线性二分可分三种情况来组织最小包含最大排除球目标函数、普通核函数、局部线性核函数, 从而实现对不同的数据分布采用不同的支持向量机分类决策算法, 提高算法的分类性能.

关键词: 支持向量机; 数据分布; 分段; 组合; 分类性能

中图分类号: TP389.1

文献标识码: A

文章编号: 1000-7180(2015)03-0094-06

Research on Piecewise Combination Support Vector Machine Based on Data Distribution Rules

CHEN Peng¹, LIU Shuang², ZUO Li³, LI Xi-zuo²

(1 College of Computer Science & Engineering, Dalian Nationalities University, Dalian 116600, China; 2 Department of Computer Science & Technology, Dalian Neusoft University of Information, Dalian 116023, China; 3 College of Economics and Management, Dalian Nationalities University, Dalian 116600, China)

Abstract: Based on analysis of data distribution rules, piecewise combination support vector machine is put forward in this paper. First, statistical analysis is adopted for the original data. Then, k -means cluster algorithm is introduced to compute cluster centers for the data. For three cases of data distribution, which is, sphere-distribution, linearly binary separable distribution, and locally linearly separable distribution, minimum enclosing and maximum excluding support vector machine, common support vector machine and local linear support vector machine are defined to solve the objective problems. Different support vector machines for different data distribution will improve classification performance.

Key words: support vector machine; data distribution; piecewise; combination; classification performance

1 引言

支持向量机算法作为机器学习领域替代神经网络的新型机器学习算法, 已经应用到图形图像处理、数据挖掘、文本分类等众多领域^[1-3]. 支持向量机算法以二分类为基础, 其基本思想是寻找一个满足分类要求的超平面, 使该超平面不但能将两类数据正

确分开, 而且使分类间隔最大, 即两类数据点距离分类面尽可能远. 通过数学理论推导目标函数最终转化成为一个二次优化问题, 从理论上说, 得到的将是全局最优解. 解决了神经网络方法中无法避免的局部极值问题. 而且, 对非线性可分问题通过非线性变换转换到高维的特征空间, 在高维空间中构造线性判别函数来实现原空间中的非线性判别函数, 可以

收稿日期: 2014-05-26; 修回日期: 2014-07-18

基金项目: 国家自然科学基金资助项目(71303031)

巧妙地解决维数问题,其算法复杂度与样本维数无关。

针对不同的应用背景,很多研究人员对二分类支持向量机进行了改进或扩展^[4-6]。在大部分实际应用中采用的支持向量机目标函数单一,极少考虑数据分布规律,统一采用一种目标函数实现。但很多实际应用的数据分布规律非常复杂,采用单一的支持向量机分类决策可能会增加目标函数求解代价,而且把训练集合中的大部分样本点都归类为支持向量,导致测试样本的分类结果计算需要和所有的支持向量进行计算,不仅增加计算量,而且分类精度不一定满足要求。

本文提出一种分段组合支持向量机算法,充分考虑数据分布规律,采用不同的决策规则来改进算法性能。

2 数据分布规律分析

本文只讨论二分类支持向量机,不考虑多分类情况,多分类情况可以采用组合二分类支持向量机的形式求解。根据实际应用数据集合的统计分布规律,分为球形分布、线性二分可分、局部线性二分可分三种情况讨论。

(1) 球形分布

数据集合符合球形分布的特点:两类数据集合聚类中心点相近;每一类数据近似符合均匀对称分布;一类数据近似包含在另一类数据内部。

用数学统计工具进行统计时,首先对每一类数据坐标进行排序,采用2均值聚类求出聚类中心,统计聚类中心是否相近;根据聚类中心计算每类数据四个象限方向的数据个数即可得到是否符合对称均匀分布;按坐标排序结果可知两类数据的坐标范围,可用于包含关系的判断。

图1所示为在Matlab中模拟的一个特殊情况,两个类别中心相同,以半径为1和半径为2分别分散数据。

因为该数据集合非线性可分,因此采用RBF核函数进行非线性转换,求得分类面如图2所示。包围内部类别的单位圆保留不动,另外一个非均匀分布的曲线就是分类超平面。此时,如果采用最小包含最大排除球目标函数定义,求解出超球分类平面的结果会更精确,而且计算代价较小。

(2) 局部线性可分

数据集合局部线性可分的特点:整个数据集合无固定的分布规律,但是划定区间内的两类数据线性可分。

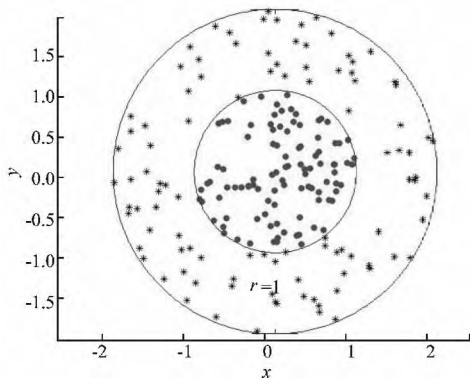


图1 球形分布模拟例子

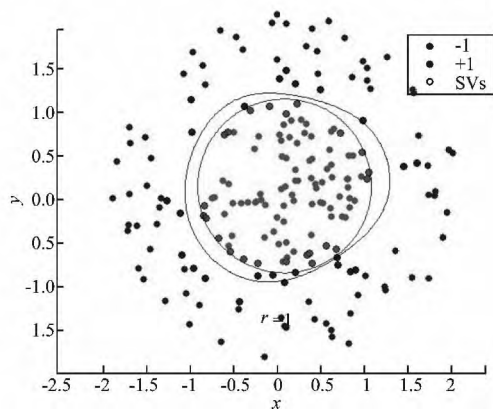


图2 图1模拟数据数据RBF核函数求解的分类结果性可分。

用数学统计工具进行统计时,首先对数据按坐标排序,然后采用 k 均值聚类求出聚类中心,考察相邻聚类中心的两类数据是否线性可分。

图3是一个做图像分类时的二类数据分类,圆圈代表一类数据,加号代表另外一类数据,图中的曲线为用支持向量机算法求得的决策曲线。我们发现,该数据集合数据很复杂,决策分类平面也很复杂。对样本分类时计算量很大,因为需要和所有的支持向量进行计算。

同时,经过对数据排序分析可知,在图3所示的各个局部区域,两类数据集实际上是线性可分的,如图4所示。因此可以采用局部线性支持向量机组合的方式求解这个复杂的决策问题,特别是对新样本的分类,只需按样本所在位置即可用对应的线性支持向量机进行分类,计算量大大降低。

为了使分析过程更直观,这里以二维数据为例,图5所示为用Matlab2014实现的模拟数据。直接用RBF核函数求解的分类决策平面如图6所示,支持向量加了一个圆圈特殊标记,由计算结果可知,不但

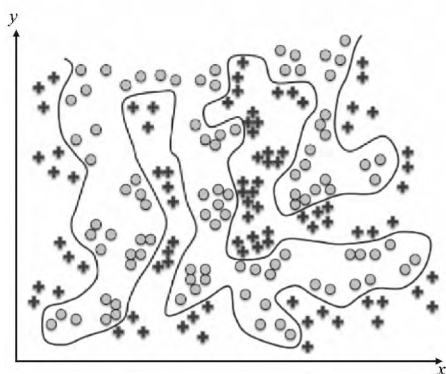


图3 支持向量机分类举例

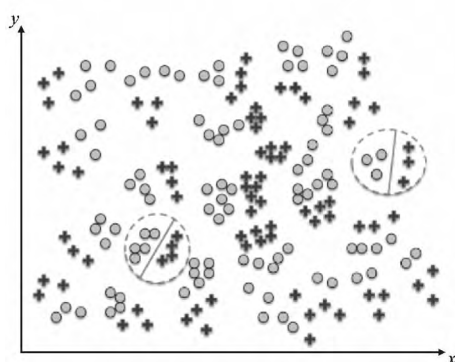


图4 局部线性可分举例

决策超平面非常复杂,而且大部分原始数据点都标记为支持向量,新样本的分类求解时要和所有的支持向量进行计算,计算代价很大.

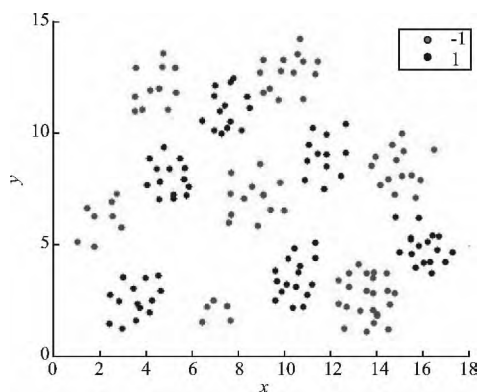


图5 Matlab 模拟产生的两类二维数据

(3) 线性可分

线性可分情况是数据分布规律最简单的一种情况.假定大小为 n 的训练样本集 $\{(x_i, y_i), i = 1, 2, \dots, n\}$ 由二类别组成,如果 $x_i \in R^d$ 属于第1类,则标记为正($y_i = +1$);如果属于第2类,则标记为负($y_i = -1$).如果存在分类超平面 $w \cdot x + b = 0$ 使得

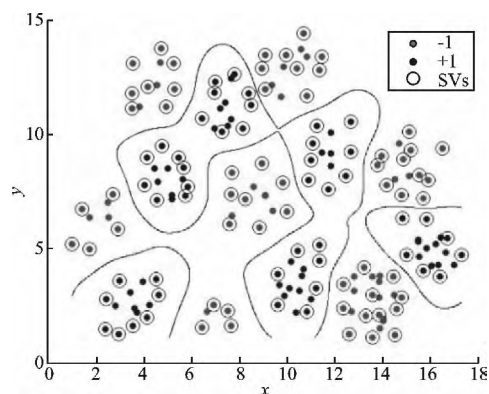


图6 模拟数据支持向量机决策平面示意图

$$\begin{cases} w \cdot x_i + b \geq +1, y_i = +1 \\ w \cdot x_i + b \leq -1, y_i = -1 \end{cases}, i = 1, 2, \dots, n \quad (1)$$

则称训练集是线性可分的,否则称训练集是线性不可分的.

对分类线方程 $x \cdot w + b = 0$ 进行归一化,使得对线性可分的样本集 $(x_i, y_i), i = 1, \dots, n, x \in R^d, y \in \{+1, -1\}$,满足:

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, \dots, n \quad (2)$$

任一点 x_i 到超平面 (w, b) 的距离为

$$d(w, b; x_i) = \frac{|(w \cdot x_i) + b|}{\|w\|} \quad (3)$$

式(2)等价于 $\min_i |(w \cdot x_i) + b| = 1$,则由(3)式可知:

$$d(w, b; x_i) \geq \frac{1}{\|w\|} \quad (4)$$

分类间隔 ρ 的表达式为

$$\begin{aligned} \rho(w, b) &= \min_{x_i: y_i = -1} d(w, b; x_i) + \min_{x_i: y_i = +1} d(w, b; x_i) \\ &= \min_{x_i: y_i = -1} \frac{|(w \cdot x_i) + b|}{\|w\|} + \min_{x_i: y_i = +1} \frac{|(w \cdot x_i) + b|}{\|w\|} \\ &= \frac{1}{\|w\|} \left[\min_{x_i: y_i = -1} |(w \cdot x_i) + b| + \min_{x_i: y_i = +1} ((w \cdot x_i) + b) \right] \\ &= \frac{2}{\|w\|} \end{aligned} \quad (5)$$

即此时分类间隔等于 $2/\|w\|$,使间隔最大等价于 $\|w\|^2$ 最小,满足式(2)且使 $\frac{1}{2} \|w\|^2$ 最小的分类面就是最优分类面,分类面上的训练样本点即为支持向量.

使分类间隔最大实际就是对推广能力的控制,这是 SVM 的核心思想之一. 统计学习理论指出,在 N 维空间中,设样本分布在一个半径为 R 的超球范围内,则满足条件 $\|w\| \leq A$ 的正则超平面构成的指示函数集

$f(x, w, b) = \text{sgn}\{(w \cdot x) + b\}$ 的 VC 维满足下面的界

$$h \leq \min([R^2 A^2], N) + 1 \quad (6)$$

因此使 $\|w\|^2$ 最小就是使 VC 维的上界最小,从而实现 SRM 准则中对函数复杂性的选择.

采用 Lagrange 乘子法求解这个具有线性约束的二次规划问题,即

$$\begin{aligned} \min \quad & \frac{1}{2} \|w\|^2 \\ \text{s. t. } & y_i[(w, x_i) + b] \geq 1 \\ & i = 1, \dots, n \end{aligned} \quad (7)$$

对应的拉格朗日多项式为

$$L_P = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i((w, x_i) + b) - 1] \quad (8)$$

式中, $\alpha_i \geq 0$ 为拉格朗日乘子. 由此得到

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^n \alpha_i y_i x_i = 0 \quad (9)$$

$$\frac{\partial L}{\partial b} = - \sum_{i=1}^n \alpha_i y_i = 0 \quad (10)$$

把式(9)、(10) 带入式(8),可以得到上述最优分类面问题的对偶优化问题,即在约束条件:

$$\sum_{i=1}^n y_i \alpha_i = 0, \alpha_i \geq 0, i = 1, \dots, n \quad (11)$$

下对 α_i 求解下列函数的最大值:

$$Q(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \quad (12)$$

式中, α_i 为与每个样本对应的 Lagrange 乘子. 这是一个不等式约束下二次函数寻优的问题,存在唯一解. 容易证明,解中将只有一部分 α_i 不为零,对应的向量就是支持向量. 解上述问题后得到的最优分类函数是:

$$\begin{aligned} f(x) &= \text{sgn}\{(w^* \cdot x_i) + b^*\} \\ &= \text{sgn}\left\{\sum_{i=1}^n \alpha_i^* y_i (x_i \cdot x) + b^*\right\} \end{aligned} \quad (13)$$

由于非支持向量的系数为 0, 式中求和实际上只对支持向量进行. b^* 是分类阈值,可以由任意一个支持向量用式(2) 求得. 为了计算可靠,也可以对所有支持向量分别计算 b 的值,然后求其平均.

3 分段组合支持向量机数学描述

上一小节已给出线性可分问题的数学描述及求解推导. 这里只给出局部线性可分问题和球形分布两种情况的数学描述.

局部线性可分问题目标函数由多个线性可分函数组成. 假设有 M 个二分支支持向量机,其目标函数定义如下:

$$\begin{aligned} \min \quad & \frac{1}{2} \|w_1\|^2 + C_1 \sum_{i=1}^{n_1} \xi_{1i}, \dots, \min \frac{1}{2} \|w_M\|^2 + \\ & C_M \sum_{i=1}^{n_M} \xi_{Mi} \end{aligned} \quad (14)$$

$$\text{s. t. } y_i F(x_i) - 1 + \xi_i \geq 0, \xi_i \geq 0, i = 1, \dots, n$$

其中,

$$F(x) = \text{sign}(f(x)) \quad X = \{(x_i, y_i)\}, i = 1, \dots, n;$$

$$f_k(x) = w_k^T \cdot x + b_k, k = 1, \dots, M.$$

式中, C 为惩罚因子,控制对错分样本惩罚的程度. ξ 为松弛因子,允许存在错分情况.

其对偶目标函数的推导过程与线性可分情况类似. 这里不再一一细致推导.

球形分布情况的目标函数^[7] 定义如下:

$$\begin{aligned} \min_{a, R^2, \Delta R^2, \xi_i} \quad & rR^2 - \Delta R^2 + C \sum_{i=1}^n \xi_i \\ \text{s. t. } & \|a - x_i\|^2 \leq R^2 - \Delta R^2 + \xi_i, y_i = +1 \\ & \|a - x_i\|^2 \geq R^2 + \Delta R^2 - \xi_i, y_i = -1 \\ & \xi_i \geq 0 \end{aligned} \quad (15)$$

式中, $0 \leq r \leq 1$ 为可调节参数, C 为惩罚因子, $\xi_i \geq 0$ 为松弛因子, $\Delta R^2 \geq 0$ 为约束条件. 引入拉格朗日乘子,求解其对偶的目标函数为

$$\begin{aligned} \max_{\alpha_i, \beta} \quad & -\frac{1}{r} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^n \alpha_i y_i x_i^T x_i \\ \text{s. t. } & \sum_{i=1}^n \alpha_i y_i = r, \sum_{i=1}^n \alpha_i - \beta = 1, \\ & \beta \geq 0, 0 \leq \alpha_i \leq C, i = 1, \dots, n \end{aligned} \quad (16)$$

4 分段组合支持向量机算法及实验

4.1 算法描述

本文算法的详细描述.

Step1 对所有数据按坐标排序.

Step2 采用 k 均值聚类算法求出 k 个聚类

中心.

Step3 如果 $k = 2$, 调用线性可分支持向量机求解最优超平面, 计算该最优平面对测试数据的分类精度, 如果精度符合要求, 算法结束, 否则转 Step4. 如果 $k > 2$, 转 Step5.

Step4 $k = 2$, 且满足两个聚类中心距离小于设定的阈值, 两个类别四个象限方向的数据个数均匀分布, 一个类别的数据近似包含在另一个类别数据内部, 调用最小包含最大排除球支持向量机算法求解最优分类超平面. 如果该最优超球面对测试数据(训练数据的一部分)的分类精度符合要求, 转 Step7, 否则对这两个子集的数据重新进行 2 均值聚类计算, 转 Step3.

Step5 对 k 个聚类中心, 重新做类别标记. 如第一个类别有五个聚类中心, 则分别标记为 $y_{11}, y_{12}, \dots, y_{15}$, 同理第二个类别的类别标记为 $y_{21}, y_{22}, \dots, y_{25}$.

Step6 考察每个聚类中心与其相邻聚类中心的类别是否相异, 如果相异调用线性可分支持向量机求解其支持向量的拉格朗日乘子; 如果相同则不做任何操作.

Step7 对新的样本或测试样本, 计算其与 k 个聚类中心的欧氏距离, 调用欧氏距离最近的几个点的分段组合决策目标函数, 取目标函数最大值作为新样本的分类结果.

Step8 如果对测试样本的分类精度符合要求, 则算法结束. 否则取距离最近的两个聚类中心的数据子集, 转 Step4 重新计算.

4.2 实验描述

为了验证本文算法的有效性, 取两组人工数据、一组实际数据进行验证. 第一组数据为了实验结果的直观性, 直接用 Matlab 产生模拟数据, 如图 7 所示, 其中, 星号代表第一类数据, 实心点代表第二类数据.

由图 7 可知, 这两类数据不是线性可分的, 如果想求解分类平面, 可以用 RBF 核函数的方式, 只是计算的代价较高. 采用本文的算法先计算聚类中心, 结果如图 8 所示, 之后对每两个相邻相异的类别计算, 就可以计算出直线的分类超平面, 计算代价较小.

第二组数据相对第一组模拟数据复杂一些, 虽然只有两个类别的数据, 但是共有 13 个子集. 原始数据如图 9 所示. 聚类中心点用叉号特殊标注, 如图 10 所示. 每两个相邻的聚类, 如果类别相异, 就

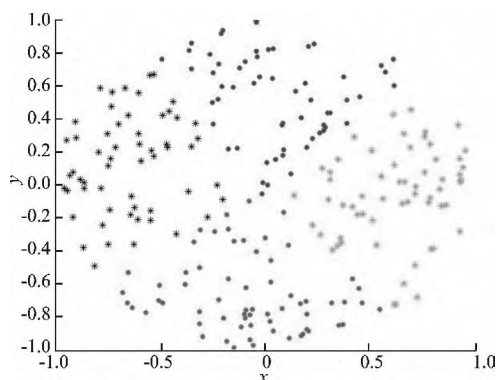


图 7 模拟数据集合 1

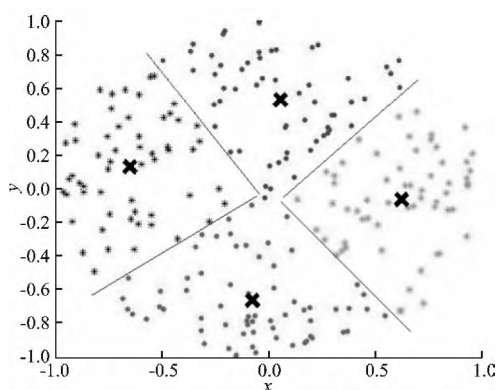


图 8 聚类中心线性分类面计算结果

求解其线性可分最优超平面, 这里就是直线. 由于相邻类别都用直线作为分类面, 支持向量个数也相对较少. 对新的测试样本, 首先计算其坐标范围, 判断其属于哪个聚类, 调用相应的决策函数计算其类别, 仅需很少的几个支持向量计算, 代价较低.

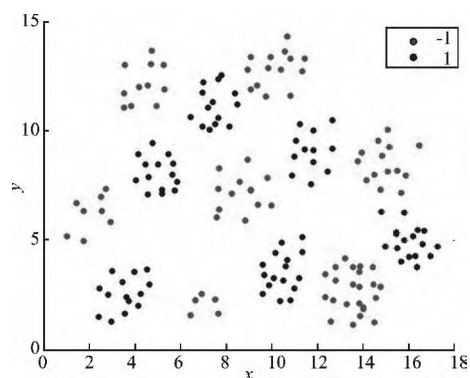


图 9 模拟数据集合 2

第三组数据为图像分类数据. 采集的动物图像数据集合大小为 5 000 幅图片, 随机选取 85% 用于训练, 15% 用于测试. 分类的过程首先计算图像的方向梯度直方图 HOG 特征, 其中两个图片的计算结果如图 11 所示, 然后调用本文算法统计数据规

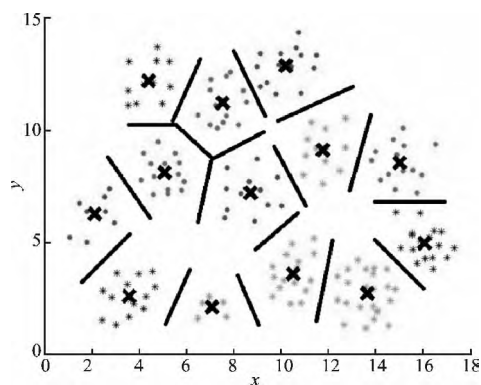


图10 聚类中心及分段线性分类面计算结果

律,选择适当的分段组合支持向量机算法求解。表1为本文算法与单一的线性支持向量机、RBF核函数支持向量机的对比运算结果,同一方法实验5次,表中数据为平均值。

表1 不同分类器的平均分类精度

方法	训练时间/s	分类精度/%	支持向量个数/个
线性 SVM	3 526	48.36	1 245
RBF 非线性 SVM	4 862	75.96	3 269
本文算法	3 923	83.45	2 312

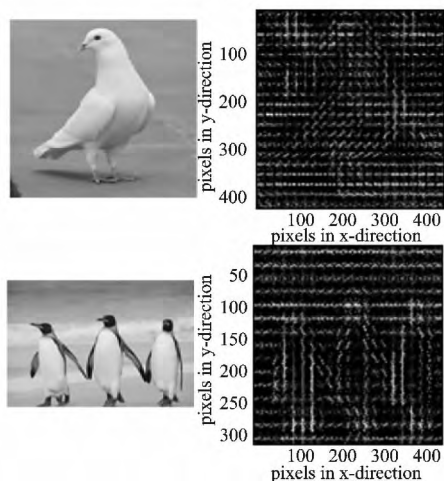


图11 图片 HOG 特征计算结果示例

由表1中的数据可知,对相同的数据集,由于本文算法采用分段组合支持向量机,考虑数据集的局部线性可分性,因此支持向量机数量小于标准的RBF非线性支持向量机算法,训练时间比线性支持向量机要长,但是测试精度较高。线性支持向量机虽然训练时间最短,但是分类精度也最低,因此不能采用简单的线性支持向量机求解。

5 结束语

本文首先分析了二分类问题的数据集分布规律,对球形分布数据集,可以采用最小包含最大排除球支持向量机求解,确保求得最优分类超平面;对线性二分可分数据,直接采用线性支持向量机求解;对局部线性二分可分数据,先计算聚类中心,然后计算相邻异类聚类中心的分类超平面,对测试样本按其坐标范围选择合适的支持向量进行分类计算,降低计算的复杂度,提高分类精度。实验结果验证了算法的有效性及可行性。尽管本文算法分析了三种情况的数据分布规律,但由于实际问题可能更复杂,因此仍需深入探讨总结数据分布规律,采用不同的支持向量机算法求解。

参考文献:

- [1] 徐冉冉, 据昊霖, 李朝锋. 非平衡二叉树主动学习支持向量机[J]. 微电子学与计算机, 2013, 30(5): 55-58.
- [2] 于静, 王辉. 改进支持向量机在网络入侵检测中的应用[J]. 微电子学与计算机, 2012, 29(3): 10-13.
- [3] 彭辉. 小波变换和支持向量机相融合的 ECG 身份识别算法[J]. 微电子学与计算机, 2013, 30(3): 152-155.
- [4] Lee G H, J S Taur, C W Tao. A robust fuzzy support vector machine for two-class pattern recognition [J]. International Journal of Fuzzy Systems, 2006, 8(2): 76-86.
- [5] 薛贞霞, 刘三阳, 刘万里. 基于 SVDD 的渐进直推式支持向量机学习算法 [J]. 模式识别与人工智能, 2008, 21(6): 721-727.
- [6] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers[J]. Neural Processing Letter, 1999, 9(3): 293 - 300.
- [7] Yi Liu, Yuan F. Zheng. Minimum enclosing and maximum excluding machine for pattern description and discrimination [C]// Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06). Hong Kong: IEEE, 2006: 129-132.

作者简介:

陈 鹏 男, (1977-), 硕士研究生, 副教授. 研究方向为计算机应用. Email: chenpeng@neusoft.edu.cn