

最小二乘支持向量机的一种非均衡数据分类算法

赵 会, 黄景涛, 谈书才

(河南科技大学 电子信息工程学院, 河南 洛阳 471003)

摘 要: 为了提高支持向量机的非平衡数据分类能力, 分析了最小二乘支持向量机的本质特征, 提出了一种基于中心距离比的非平衡数据分类算法, 同时通过修剪边界样本, 解决了最小二乘支持向量机缺失稀疏性的问题. 在 UCI 标准数据集上进行的试验表明: 该算法能够有效地提高支持向量机对非均衡分布数据的正确性, 且该算法在不影响训练精度的前提下, 可以得到稀疏解, 算法的训练速度也有了一定的提高.

关键词: 最小二乘支持向量机; 非均衡数据分类; 稀疏性

中图分类号: TP181

文献标识码: A

文章编号: 1000-7180(2010)04-0033-05

Research on Classification of Imbalanced Data Based on Least Squares Support Vector Machine

ZHAO Hui, HUANG Jing-tao, TAN Shu-cai

(Electronic Information Engineering College, Henan University of Science & Technology, Luoyang 471003, China)

Abstract: To improve the classification performance of unbalanced datasets, the nature characteristics of sparse least squares support vector machines (LS-SVM) was analyzed and an algorithm based on center distance ratio for the unbalanced samples was proposed. Meanwhile, the problem of sparseness lacking in the least squares support vector machines was solved by pruning the boundary samples. The new algorithm was tested on the UCI datasets. The results indicate that this method can effectively improve the classification accuracy of LS-SVM for the unbalanced samples, the proposed algorithm can properly obtain the sparse solutions without affecting the capacity of classification, and the speed of classifiers is also improved.

Key words: least squares support vector machines; unbalanced data classification; sparseness

1 引言

支持向量机 (Support Vector Machines, SVM) 是建立在统计学习理论之上的一种新的机器学习方法^[1], 近年来已被成功地应用在模式识别、函数回归、预测模型、图像处理等领域.

对 SVM 分类时错分的样本进行分析时发现, 支持向量机对平衡数据集有良好的分类能力. 它利用接近边界的少数向量来构造一个最优分类面, 通过最优分类面对数据进行分类. 然而对非平衡数据集, 分类面会向样本数量较少的一类移动, 从而使支

持向量机过拟合样本点数目多的一类, 而低估样本点数目少的一类, 因此对于非平衡数据集, 支持向量机的分类能力会有明显的下降.

经典的 SVM 的训练算法需要求解一个凸二次规划问题^[2], 计算的复杂性较大, 为减小支持向量机的计算复杂度, Suykens 等人^[3]1999 年提出了最小二乘支持向量 (Least Squares Support Vector Machine, LS-SVM), 该算法只需通过求解一组线性方程而获得最优分类面, 学习速度较快, 而且内存需求少, 但最小二乘支持向量机对支持向量失去了稀疏性.

文中运用 LS-SVM 针对 SVM 非均衡数据分类

收稿日期: 2009-01-30; 修回日期: 2009-03-30

基金项目: 国家自然科学基金项目 (70671035); 河南省重点科技攻关项目 (082102210015)

及 LS-SVM 缺失稀疏性的问题,利用中心距离比、支持向量、增量学习及边界向量之间的关系,提出一种基于 LS-SVM 增量学习的非均衡数据分类算法。

2 最小二乘支持向量机原理

类似于传统 SVM,LS-SVM 也是通过构造最优分类超平面实现分类的.用于分类的最小二乘支持向量机算法推导如下:设独立同分布线性不可分样本集为

$$T = (x_k, y_k), x_k \in R^m, y \in \{+1, -1\} k = 1, \dots, N$$

根据结构风险最小化原则,最小二乘支持向量机约束优化问题可表示为

$$\min_{w, b, e} J_p(w, e) = \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{k=1}^N e_k^2 \quad (1)$$

$$\text{s. t. } y_k [w^T \varphi(x_k) + b] = 1 - e_k, k = 1, \dots, N \quad (2)$$

为求解此优化问题,定义 Lagrangian 函数:

$$L(w, b, e, \alpha) = J_p(w, e) - \sum_{k=1}^N \alpha_k \{y_k [w^T \varphi(x_k) + b] - 1 + e_k\} \quad (3)$$

根据 KKT(Karush-Kuhn-Tucker) 条件有

$$\begin{cases} \frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^N \alpha_k y_k \varphi(x_k) \\ \frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^N \alpha_k y_k = 0 \\ \frac{\partial L}{\partial e_k} = 0 \rightarrow \alpha_k = \gamma e_k \\ \frac{\partial L}{\partial \alpha_k} = 0 \rightarrow y_k [w^T \varphi(x_k) + b] - 1 + e_k = 0 \end{cases} \quad (4)$$

从方程组(4)中消去 w 和 e_k 后,可以得到

$$\begin{bmatrix} 0 - Y^T \\ Y \Omega + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix} \quad (5)$$

式中, $Y = [y_1, y_2, \dots, y_N]$, $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_N]$, $\Omega_{ij} = y_i y_j K(x_i, x_j)$, $I = [1, 1, \dots, 1]$

所以,式(1)、式(2)的分类问题通过求解式(3)~式(5)的线性方程获得,而不用解决二次规划问题.通过求解上面的方程组,就可以求得如下的最优分类器:

$$y(x) = \text{sign} \left[\sum_{k=1}^N \alpha_k y_k K(x, x_k) + b \right] \quad (6)$$

从上面可以看出,最小二乘支持向量机的解中,

所有的 α_k 都不为零,失去了解的稀疏性^[4].

3 非平衡数据集分类问题的研究现状

文献[5]中指出,当两类训练样本数目相差悬殊时,分类平面明显偏向少数类训练样本一方,从而产生“有偏性”.为了解决该问题,国内外学者提出了多种解决方法,这些方法主要通过数据预处理和改造分类算法来解决非平衡数据集分类问题.

用数据预处理方法进行非平衡数据集分类是先对训练数据集进行预处理,然后用处理过的训练数据集训练分类器.数据预处理的方法包括向下取样、向上取样.向下取样是减少多数样本的数量,使不同类的数量接近平衡,从而提高分类能力.该方法的缺点是容易丢失多数类的一些重要信息.向上取样是通过增加少数类的样本数量提高分类器的分类能力.但若是简单的复制少数样本来增加少数类的数量,并不会为训练集增加有用信息,反而会使决策域变小,导致过学习.同时,人为地增加样本有可能导致增加噪点,从而降低分类精度.

改造分类算法主要通过通过对不同类的样本设置不同的权值,改变概率密度,调整分类边界等措施解决.例如,文献[6]中对 C-SVM 支持向量机进行改造,对不同的类设置相同的惩罚参数 C ,如果两类别样本数目不等,则意味着哪一类样本点的个数多,就更看重哪类点,但是实际希望对正类点和负类点的惩罚是相同的,因此,该方法对非平衡数据集分类效果是不明显的.

4 基于 LS-SVM 增量学习的不均衡数据分类算法

通过对目前非平衡分布数据分类的方法研究,文中基于最小二乘支持向量机稀疏化和增量学习算法,提出了一种新的针对不平衡数据的分类算法 SILS-SVM (sparse imbalance LS-SVM).

该算法的主要思想:将 4.3 节中选出的边界样本加上剩余的训练样本中的那些错分点组成新的训练集,再用最小二乘支持向量机算法重新训练,得到新的分类面,即采用增量学习和最小二乘支持向量机稀疏化算法相互交替的办法进行训练,在每次训练时保持两类样本数的基本平衡.

4.1 增量学习算法

文中利用增量学习的思想^[7]对其进一步改进,算法思路描述如下:首先选择训练样本,文中选择那些靠近边界的样本,因为只有这些样本才可能成为

支持向量(选择的目的是为了进一步减少训练样本数,从而提高算法的速度),然后利用边界样本进行SVM的增量学习,求出支持向量,最终求出最优分类面.文中增量学习的目的是为了避开支持向量选取不全的情形,保证了训练样本包括所有的支持向量所对应的训练样本,从而在保证了支持向量机的分类能力没有受到影响的前提下,提高了支持向量机的训练速度.

4.2 稀疏性最小二乘支持向量机算法

实验表明仅使用增量学习,工作集中多数类样本的支持向量数目将会越来越多,这将引起训练集的再一次不平衡.对于最小二乘支持向量机分类问题,只有少量被称为支持向量的训练样本才对分类起作用,即使删除其他所有训练样本,仍然可以使用剩余的支持向量子集找到相同的最大间隔超平面^[7].所以在提出的算法中,每次增量学习之后,马上进行最小二乘支持向量机稀疏化.通过该方法,减少工作集中多数类样本的支持向量数目.

4.3 基于LS-SVM增量学习的不均衡数据分类算法

对于分类问题,边界样本包含了分类的重要信息,最难分类的和最容易引起分类错误的样本也是集中在边界附近,如果失去了训练样本的边界信息,必然会使得SVM不能更好的体现样本的分布信息.

由文献[8]样本所对应的样本点是对应中心距离比较大的样本点的原则,且边界向量样本所对应的训练样本仅占总训练样本的一小部分($\leq 10\%$),且支持向量样本包含在边界向量样本集合内,因此通过设置阈值选取中心距离比较大的(该比率可以根据实际情况调节)的训练样本点作为边界向量样本集合.这样既解决了阈值选取的难点,使中心距离比也得到了有效的利用.

依据上面的分析,对每类样本中的所有样本点求出该样本点对应的中心距离比,然后将该类中的中心距离比进行排序,取中心距离比值较大(约20%)所对应的训练样本点作为该类的边界向量样本集合;另一类也同样如此,即可抽取出包含支持向量所对应训练样本在内的边界向量样本集合,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.

定义1(距离) 给定两个训练样本 x_i, x_j , 则两样本点在再生核 Hilbert 空间的距离可表示为

$$\begin{aligned} d(\phi(x_i), \phi(x_j)) &= \\ &= \sqrt{K(x_i, x_i) - 2K(x_i, x_j) + K(x_j, x_j)} \\ &= \|\phi(x_i) - \phi(x_j)\|^2 \end{aligned} \quad (7)$$

定义2(中心距离) 就是样本到中心的距离.假设两类训练样本数分别为 N_+, N_- . 则两类的类

中心^[5] 分别为 $\phi(X_c^+) = \frac{1}{N_+} \sum_{i=1}^{N_+} \phi(x_i^+)$, $\phi(X_c^-) = \frac{1}{N_-} \sum_{i=1}^{N_-} \phi(x_i^-)$, 且对每个正样本来说,都有两个中心

距离:自中心距离 $D_{si}^+ = d(\phi(x_i^+), \phi(x_c^+))$ 和互中心距离 $D_{mi}^+ = d(\phi(x_i^+), \phi(x_c^-))$, 同理对于每个负类样本点的自中心和互中心距离分别为

$$D_{si}^- = d(\phi(x_i^-), \phi(x_c^-))$$

$$D_{mi}^- = d(\phi(x_i^-), \phi(x_c^+)).$$

定义3(中心距离比) 已知两类样本,某一类中样本点 x_i 的自中心距离和两类中心间的距离的值分别为 D_{si} 和 D_{mi} , 则该样本点的中心距离比: $R_i = D_{si}/D_{mi}$.

算法具体实现步骤如下:

(1) 根据定义2求出两类样本的类中心,并计算两类类中心之间的距离 $D = d(\phi(X_c^+), \phi(X_c^-))$.

(2) 分别计算各类训练样本所构成的近似平均半径.在计算之前,首先去除类中的奇异点,即那些离类中心的距离值非常大(约占该类样本点数量的

$$5\% \text{ 左右}) \text{ 的样本点, 其中 } \overline{R_+} = \frac{\sum_{k=1}^{N_+} D_{si}^+}{N_+}, \overline{R_-} = \frac{\sum_{k=1}^{N_-} D_{si}^-}{N_-}.$$

(3) 利用定义3分别求出高维空间中每类样本点的中心距离比集合 R . 若已知正类训练样本集合 $X^+ \{x_i \mid x_i \in R^m, y_i = 1, i = 1, 2, \dots, N^+\}$, 则 $R^+ = \{R_i^+ \mid R_i^+ \in D_{si}^+/D_{mi}^+, i = 1, 2, \dots, N^+\}$, 同理求 R^- .

(4) 设置阈值,选取参与训练的边界样本.因为最优分类面总是位于两类训练样本所构成的凸包之间.因此根据各类训练样本中心距离比 R_i , 结合两类类中心之间的距离 D 以及近似平均半径 $\overline{R_+}, \overline{R_-}$, 来确定用于选取中心距离比较大的边界样本的阈值.

边界样本的选取方法:当 $R_i^+ > T_+$ 和 $R_i^- > T_-$ 时,则该训练样本选定为可能中心距离比较大的样

本. 这里 T_+ 、 T_- 分别是两类样本进行选择的阈值, 可以不相等.

当 $D > \bar{R}_+ + \bar{R}_-$, T_+ 、 T_- 可以取较大值, 如 $T_+ = \bar{R}_+/D - \mu$, $T_- = \bar{R}_-/D - \mu$ ($\mu = \bar{R}/4$).

当 $D < \bar{R}_+ + \bar{R}_-$, T_+ 、 T_- 可以取较小值, 如 $T_+ = \bar{R}_+/D - \mu$, $T_- = \bar{R}_-/D - \mu$ ($\mu = \bar{R}/2$).

对均衡数据, 边界样本数可取同一个比例初值, 对于非均衡数据, 选取边界样本的比率和原始训练样本数成反比^[9], 改变 μ 的值可以改变所选取边界样本的数量, 并设 T_b 为选取的边界样本集合.

(5) 以 T_b 作为 LS-SVM 的训练样本, 利用式 (5) 求出最优解 a^* 和 b^* , 得到式 (6) 确定的分类决策函数 $y_b(x)$.

(6) 记 S 为 $T - T_b$ 中那些被 $y_b(x)$ 错分的点构成的集合, 以 $T_b \cup S$ 作为新的训练集, 用 LS-SVM 重新进行训练(增量学习), 从而得 SILS-SVM 的分类决策函数 $y(x)$.

5 仿真实验

为了验证算法的有效性, 进行了一系列实验: 即先按照上述方法根据样本的中心距离比选择边界样本, 然后采用 LS-SVM 算法包进行训练^[10].

本实验用 SILS-SVM 分类方法和 LS-SVM 方法进行训练时, 核函数采用线性核函数, γ 为 SILS-SVM 的惩罚参数. 因为边界样本一般占训练样本总数的 20%~30%, 故在 SILS-SVM 算法中, 以选取三分之一左右的训练样本为标准, 来决定阈值. 训练结果如表 1~表 3 所示. 实验平台为 PC 机 (Intel 1.6 G/1G DDR). 注: 表格中试验结果顺次均为 LS-SVM 算法、SILS-SVM 算法.

表 1 Wine 12 数据分类实验结果 ($C=10, \gamma=1, 12.3\%$)

训练样 数目	训练样本集		测试样本集	
	分类正确/	学习时间/	分类正确/	花费时间/
	%	s	%	s
30,35	99.31	0.343	99.30	0.261
7,10	99.41	0.162	98.53	0.056

表 2 Wine23 未经纠偏数据分类实验结果
($C=10, \gamma=10, 11.3\%$)

训练样 数目	训练样本集		测试样本集	
	分类正确/	学习时间/	分类正确/	花费时间/
	%	s	%	s
60,20	98.75	0.541	99.12	0.168
12,4	48.76	0.040	76.97	0.051

表 3 Wine 23 经纠偏数据分类实验结果
($C=10, \gamma=10, 11.3\%$)

训练样 数目	训练样本集		测试样本集	
	分类正确/	学习时间/	分类正确/	花费时间/
	%	s	%	s
60,20	98.75	0.541	99.12	0.168
12,12	96.27	0.063	98.78	0.051

对 wine1、wine2 的分类问题 wine12, 对该组基本均衡数据, SILS-SVM 中两类边界向量样本占训练样本的比例初值都选为 20%, 对 wine2、wine3 的分类问题 wine23, 对该组非均衡训练数据, CLS-SVM 通过设置阈值预选不同比率 (20%、60%) 的两类边界向量样本的比例初值使训练样本达到均衡, 从而对分类进行纠偏, 从实验结果可以看出, 在采用相同的核函数和相同的惩罚参数时, 与 LS-SVM 相比, SILS-SVM 算法具有如下优点: 对非均衡数据, 通过增量学习和稀疏化算法交替训练的方法, 使的训练样本集基本达到平衡, 纠正偏差后的效果更好, 且支持向量所占比例越低, 算法的优势体现得越明显. 同时算法也弥补了最小二乘支持向量机的稀疏性, 减少了存储和计算量, 训练和决策速度得以提高, 且分类器在训练样本集和测试样本集上的准确率基本没有改变.

通过以上实验分析验证了支持向量机增量学习不均衡数据分类算法的有效性和可行性: 即该方法有效的利用了中心距离比值, 且适合于增量学习, 针对分布不均匀的训练数据提出了一种选择边界向量的方案, 通过选择一部分对决策分类面起作用的边界样本的方法来减少训练样本的数量, 从而在保证算法推广能力的同时也提高了支持向量机的训练速度.

6 结束语

为了提高支持向量机对非平衡数据集的分类性能. 文中根据中心距离比和边界样本、支持向量以及增量学习之间的关系, 提出了一种基于 SVM 增量学习不均衡数据的分类算法. 该算法主要具有以下几个优点: (1) 有效地利用了中心距离比值弥补了 LS-SVM 所缺失的稀疏性; (2) 解决了 LS-SVM 方法中因训练样本的不均衡性所带来的分类面的偏移问题; (3) 适合于增量学习, 添补预抽取的边界向量样本可能包括不了全部支持向量的漏洞; (4) 在不影响 SVM 分类精度的前提下, 提高了 LS-SVM 的训练速度; 实验结果进一步印证了该方法的有效

性和可行性.

参考文献:

- [1] Vapnik V N. 统计学习理论[M]. 许建华, 张学工, 译. 北京: 电子工业出版社, 2004.
- [2] 徐勋华. 支撑向量机的多类分类方法[J]. 微电子学与计算机, 2004, 21(10): 149-152.
- [3] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers [J]. Neural Processing Letter, 1999, 9(3): 293-300.
- [4] 王海峰. 最小二乘支持向量机的一种稀疏化算法[J]. 计算机工程与应用, 2005, 43(33): 68-70.
- [5] Wu G, Chang E Y. Class-boundary alignment for imbalanced dataset learning[C]//Workshop on Learning from Imbalanced Datasets (ICML 2003). Washington, DC, 2003: 49-56.
- [6] 刘爽, 贾传炎, 陈鹏. 一种自动选择参数的加权支持向量机算法[J]. 计算机工程与应用, 2006, 42(2): 64-66.
- [7] 孔锐. 一种快速支持向量机增量学习算法[J]. 控制与决策, 2005, 20(10): 1129-1132.
- [8] 姚全珠. 基于最小二乘支持向量机的非平衡分布数据分类[J]. 计算机工程与应用, 2008, 44(5): 167-170.
- [9] Pelckmans K, Suykens J A K, Van T Gestel. et al. LS-SVMLab1. 5Toolbox[EB/OL]. (2003-10-11)[2009-07-21]. <http://www.esat.kuleuven.ac.be/sista/lssvm-lab>.
- [10] UCL Machine Learning Group. Elena database[EB/OL]. [2009-07-15]. <http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/elena.htm#stuff>.

作者简介:

赵会女, (1982-), 硕士研究生. 研究方向为模式识别与智能系统.
黄景涛 男, (1977-), 博士, 副教授. 研究方向为数据挖掘和智能优化算法及应用.
谈书才 男, (1984-), 硕士研究生. 研究方向为模式识别与智能系统.

(上接第32页)

(Persistence Tool)类, 给出了针对 JavaBean 一种简便的持久化方案. 在此基础上构建数据库应用系统, 可简化代码编写任务, 加快系统开发进程. Hibernate、Toplink 等开源框架虽然也提供了类似功能, 但都比较复杂^[5]. 使用本工具类不需要外部配置文件, 比较适合于中小型项目的开发. 至于由于引入反射技术而造成程序性能下降的问题, 可通过字节码增强技术与适当的缓存来改善, 限于篇幅不再赘述.

参考文献:

- [1] 唐姗, 赵文耘. 基于反射的动态软件体系结构实现[J]. 微电子学与计算机, 2006, 23(9): 32-37.
- [2] 吴卫平, 王丽芳, 蒋泽军. 基于 ORM 的数据持久层框架

- 研究[J]. 微电子学与计算机, 2008, 25(7): 188-193.
- [3] 张义. Java5 程序员开发手册[M]. 北京: 北京希望电子出版社, 2007.
- [4] Holzner S. Java 技术内幕[M]. 马朝晖, 译. 北京: 机械工业出版社, 2002.
- [5] 杨飞, 袁建华. Hibernate 框架在权限管理系统中的应用[J]. 微电子学与计算机, 2007, 24(2): 206-208.

作者简介:

檀明 男, (1974-), 博士研究生, 讲师. 研究方向为数据库与多媒体技术、工业控制网络.
许强 男, (1972-), 硕士, 讲师. 研究方向为软件工程、计算机网络.
李丛 男, (1985-), 助教. 研究方向为软件工程.