适用于不平衡样本数据处理的支持向量机方法

吴洪兴,彭 宇,彭喜元
(哈尔滨工业大学自动化测试与控制系,黑龙江哈尔滨 150080)

摘 要: 支持向量机算法在处理不平衡样本数据时,其分类器预测具有倾向性.样本数量多的类别,其分类误差小,而样本数量少的类别,其分类误差大.本文针对这种倾向性问题,在分析其产生原因的基础上,提出了基于遗传交叉运算的改进方法.对于小类别训练样本,利用交叉运算产生新的样本,从而补偿了因训练数据类别大小差异而造成的影响.基于 UCI 标准数据集的仿真实验结果表明,改进方法比标准支持向量机方法具有更好的分类准确率.

关键词: 支持向量机; 交叉算子; 类别差异; 模式识别
中图分类号: TP391.4 **文献标识码**: A **文章编号**: 0372-2112 (2006) 12A-2395-04

A New Support Vector Machine Method for Unbalanced Data Treatment

WU Hong-xing, PENG Yu, PENG Xi-yuan
(Department of Automatic Test and Control, Harbin Institute of Technology, Harbin, Heilongjiang 150080, China)

Abstract: In SVM algorithm, when training sets with uneven class sizes are used, the prediction result of classifier is undesirably biased towards the class with more samples in the training set. That is to say, the larger the sample size, the smaller the classification error, whereas the smaller the sample size, the larger the classification error. Aiming at this orientation problem and with the analysis of the cause of it, an improved method based on genetic crossover operator was proposed, for the training set with small size generate new samples by using crossover operation, thereby compensates for the unfavorable impact caused by the bias of the training data class size. Simulation experiment results on UCI stander data shows that the proposed method has better classification accurate compared with stander support vector machine method.

Key words: support vector machine (SVM); crossover operator; uneven class size; pattern recognition

1 引言

支持向量机是由 Vapnik 于 1992 年提出的一种新的机器学习方法[1].由于其出色的性能,已经在人脸识别、生物信息、数据挖掘、文本分类等诸多领域得到了广泛的应用.在研究过程中人们发现支持向量机在各类别样本数多少不同时,其预测过程具有倾向性,对样本数量多的类别,其训练误差和预测误差小;而对样本数量少的类别,其训练误差和预测误差大.Chew Hong-Gunn 在文献[2]中详细分析了 C-SVM 算法中因类别大小不平衡而造成对分类精度影响的原因,并提出了相应的解决方法.在后续研究中 Chew 又提出了用双 ν-SVM 算法[3]来解决 ν-SVM 算法[4]训练类别大小不均衡带来的问题,文献[5]提出了一种模糊支持向量机算法,该算法可以减少外部和数据中噪声的影响.在文献[6]中范昕炜等提出了加权支持向量机算法,补偿了支持向量机这种预测倾向性造成的不利影响.

针对在不同类别样本不平衡的情况下,支持向量机的预测具有倾向性的缺陷,本文在说明预测倾向性现象产生原因的基础上,提出改进的支持向量机算法.对于样本数量较少的类别,利用遗传算法中的交叉算子对样本进行处理,构造出符合原样本分布要求的新样本.使不同类别的样本数量趋于平衡,从而在支持向量机训练过程中对类别差异造成的影响进行相应的补偿,提高整体的分类精度,这对于某些需要重点关注小类别精度的应用场合有重要的现实意义.

2 不平衡样本数据对分类精度的影响

2.1 支持向量机算法的基本原理

SVM 方法是从线性可分情况下的最优分类面提出的.最优分类面是一种分类超平面,它不但能够将所有训练样本正确分类,而且使训练样本中离分类面最近的点到分类面的距离(定义为间隔)最大,通过使间隔最大化来控制分类器的复杂度,进而实现较好的泛化能力.在两类模式识别问题中,给定训练数据 (x_i, y_i), $i = 1, \cdots, n$, $x \in R^d$, $y \in \{+1, -1\}$,支持向量机就是指由下式确定的分类规则:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \quad (1)$$

收稿日期:2006-08-22;修回日期:2006-11-13

式中的 $\alpha_i \geq 0$, $i = 1, \cdots, n$ 是下面优化问题的最优解:

$$\begin{aligned} &\min \Phi(w, \xi) = \frac{1}{2} \| w \|^2 + C\left(\sum_{i=1}^{l} \xi_i\right) \\ &y_i(x_i \cdot w + b) \geq 1 - \xi_i, \\ &i = 1, \cdots, n \quad \xi_i \geq 0 \end{aligned} \tag{2}$$

利用 Lagrange 优化方法可以把上述最优化问题转化为其对偶问题:

$$\begin{aligned} &\text{Max} \quad W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j}^{l} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ &\text{s.t} \quad \sum_{i=1}^{l} \alpha_i y_i = 0 \\ &0 \leq \alpha_i \leq C \quad i = 1, \cdots, l \end{aligned} \tag{3}$$

上述各式中, $w = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)$ 为与分类间隔有关的变换向量; ξ_i 为松弛因子; C 为惩罚参数; α_i 为拉格朗日乘子, α_i 不为零的样本就是支持向量; $k(\cdot)$ 是满足 mercer 条件的核函数,其作用是将输入空间映射到高维的特征空间,然后在特征空间中寻找上述优化问题的最优解. 常用的核函数主要有多项式核函数、高斯核函数、Sigmoid 函数,本文采用径向基核函数(高斯核函数).

2.2 不平衡样本数据对分类精度的影响

训练好的支持向量机间隙宽度在大多数情况下是不为零的. 在最优化求解式(3)得到的拉格朗日系数 α_i 可能会有三种情况:

(1) $\alpha_i = 0$,这时 x_i 被正确分类;

(2) $0 \leq \alpha_i \leq C$,此时所对应的 x_i 称为标准支持向量(Normal Support Vector),标准支持向量是两类样本中离分离面最近,且平行于最优超平面的训练样本;

(3) $\alpha_i = C$,此时所对应的 x_i 称为边界支持向量(Boundary Support Vector),其实际是错分的样本点,边界支持向量的比例反映了支持向量机的分类准确率.

假设 N_{BSV+} 和 N_{BSV-} 分别代表正类和负类中边界支持向量的个数, N_{SV+} 和 N_{SV-} 分别代表正类和负类中所有支持向量的个数, $M+$ 和 $M-$ 分别代表正类和负类中的样本数. 根据式(3)有:

$$\sum_{i=1}^{l} \alpha_i = \sum_{y_i=+1} \alpha_i + \sum_{y_i=-1} \alpha_i \tag{4}$$

$$\sum_{y_i=+1} \alpha_i = \sum_{y_i=-1} \alpha_i \tag{5}$$

因为所有 α_i 的最大值是 C,所以有:

$$N_{BSV+} \times C \leq \sum_{y_i=+1} \alpha_i \tag{6}$$

$$N_{SV+} \times C \geq \sum_{y_i=+1} \alpha_i \tag{7}$$

将式(6)与式(7)结合,得到:

$$N_{BSV+} \times C \leq \sum_{y_i=+1} \alpha_i \leq N_{SV+} \times C \tag{8}$$

类似的,可以得到:

$$N_{BSV-} \times C \leq \sum_{y_i=-1} \alpha_i \leq N_{SV-} \times C \tag{9}$$

设 $\sum_{y_i=+1} \alpha_i = \sum_{y_i=-1} \alpha_i = L$,用式(8)和式(9)分别除以 $C \times M+$ 和 $C \times M-$,得到:

$$\frac{N_{BSV+}}{M+} \leq \frac{L}{C \times M+} \leq \frac{N_{SV+}}{M+} \tag{10}$$

$$\frac{N_{BSV-}}{M-} \leq \frac{L}{C \times M-} \leq \frac{N_{SV-}}{M-} \tag{11}$$

由式(10)和式(11)可知:如果 $M+ \neq M-$,则正类和负类中边界支持向量比例的上界和支持向量比例的下界不相等. 样本数少的类别其边界支持向量比例的上界比样本数多的类别的边界支持向量比例的上界大. 这意味着样本数少的类别中的样本被错分的比例要比样本数大的类别中的样本被错分的比例大. 例如正样本数 $M+$ 大时, $\frac{L}{C \times M+}$ 小, $\frac{N_{BSV+}}{M+}$ 小,即错误分类率小,反之亦然.

3 基于交叉算子的新样本生成方法

在支持向量机方法中,当训练集中不同类别之间样本数量趋于平衡时,预测倾向性会急剧减小. 受遗传算法交叉运算的启发,针对样本数量少的类别利用交叉运算方法生成新的样本,使不同类别的样本数量趋于平衡,从而提高预测的准确率.

遗传算法中的所谓交叉运算,是指对两个相互配对的染色体按某种方式相互交换其部分基因,从而形成两个新的个体. 交叉运算是遗传算法区别于其他进化算法的重要特征,它在遗传算法中起着关键作用,是产生新个体的主要方法[7]. 具体的方法如下:

首先,在交叉运算之前先对样本数量较少的类别中的个体进行配对. 目前常用的配对策略是随机配对,即将类别中的 m 个个体以随机的方式组成 $m/2$ 对配对样本组,交叉操作是在这些配对样本组中的两个个体之间进行的. 每一次交叉操作会生成两个新的样本,同时父代样本被保留,这样每一轮交叉运算以后,类别中会保留 $2m$ 个样本,如果类别中的样本数量还没有达到要求,可以进行下一轮交叉运算,直至获得所需数量的训练样本.

为了保障新生成的样本符合原有样本的属性特征,采用新样本到原有样本所属类别中心的距离来判断新样本是否满足要求. 以下给出本文涉及到的一些定义:

定义 1 某一类样本的平均特征称为该类样本的中心,已知样本向量组 $\{x_1, x_2, \cdots, x_n\}$,那么其中心为:

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{12}$$

定义 2 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, x_2,其样本距离为:

$$d(x_1, x_2) = \| x_1 - x_2 \|_2 = \sqrt{\sum_{i}^{N} (x_1^i - x_2^i)^2} \tag{13}$$

定义 3 中心距离指的是各样本到中心的距离.

$$d(x, m) = \sqrt{\sum_{i=1}^{N} (x^i - m^i)^2} \tag{14}$$

验证一个新样本是否符合在输入空间中原类别的分布,首先要依据式(12)计算原类别的中心,然后依据式(13)计算类别中原有样本到类别中心的距离.选取这些距离中的最大值作为阈值.然后再计算每一个新生成样本到类别中心的距离,当距离值大于所设的阈值时,将这个新生成的样本去除,当距离值小于所设的阈值时,将新样本存入训练集.重复上述过程,直到训练集中不同类别样本数量趋于平衡.

4 仿真实验

为了验证本文提出方法的有效性,从 UCI 机器学习知识库[8]中选取了 Sonar、Pima-Indians-Diabetes、Breast Cancer 三个数据集进行实验.其中 Pima-Indians-Diabetes、Breast Cancer 数据集正、负两类样本不平衡,Sonar 数据集基本均衡,为了验证本文提出的方法,在 Sonar 数据集正类样本中随机选取 20 个样本与所有负类样本组成 Sonar1 数据集;在负类样本中随即选取 20 个样本与所有正类样本组成 Sonar2 数据集.以上每个数据集中的样本随机选取 70 %用于训练,其余用于测试.所用数据集的基本信息见表 1.

表 1　数据集基本信息

数据集	正样本数	负样本数	特征数	类别数	训练样本数	测试样本数
Sonar1	20	111	60	2	92	39
Sonar2	97	20	60	2	82	35
Pima	268	500	8	2	537	231
Breast Cancer	458	241	30	2	489	210

实验是在 MATLAB 环境下进行的.实验过程中使用了 OSU-SVM MATLAB 工具箱[9]以及英国谢菲尔德(Sheffield)大学开发的基于 MATLAB 的遗传算法工具箱[10].实验过程中,支持向量机的核函数采用径向基核函数,运行参数采用交叉验证的方法选取.交叉运算采用两点交差,交叉率为 0.8.每一个实验数据集反复进行了 10 次试验,最终结果是 10 次实验结果的平均值.采用原始数据集的试验结果见表 2.利用交叉运算生成新样本组成新的训练集后,其实验结果见表 3.

表 2　原始数据集实验结果

数据集	原始样本				
	正训练样本	负训练样本	C	δ	分类准确率
Sonar1	14	78	3800	0.3	66.72 %
Sonar2	68	14	3800	0.3	82.27 %
Pima	188	350	1100	0.6	62.33 %
Breast Cancer	320	169	2000	0.01	92.55 %

表 3　本文方法实验结果

数据集	交叉运算产生新的样本集合					
	新生成样本	调整后的正训练样本	负训练样本	C	δ	分类准确率
Sonar1	60	74	78	3800	0.3	76.02 %
Sonar2	50	68	64	3800	0.3	84.58 %
Pima	160	348	350	1100	0.6	68.88 %
Breast Cancer	150	320	319	2000	0.01	94.1 %

由表 2 中的数据可以看出,这几个数据集原始训练数据正、负类别之间样本的数量相差较大.针对样本数量较少的类别,采用前面叙述的交叉运算的方法,利用原始的训练样本作为父代生成新的样本,使每一个数据集用于训练的正、负类样本的数量趋于平衡,表 3 中的数据反映了这一趋势.从表 2、3 中的实验结果可以看出,利用新生成的不同类别样本数量比较均衡的训练数据集训练支持向量机分类器,其分类准确率要高于使用原始训练数据集训练的支持向量机分类器.其中 Pima 数据集在本文设定参数的情况下分类准确率提高了 6.55 %;Breast Cancer 数据集分类准确率提高了 1.55 %;而 Sonar1 和 Sonar2 数据集分类准确率则分别提高了 10.3 %和 1.31 %.这主要是由于增加了小类别训练样本的数量,降低了所训练分类器的预测倾向性.从而提高了分类器的整体分类准确率.

5 结论

本文针对支持向量机算法因不同类别样本数量差异造成的分类器预测具有倾向性的问题,在说明造成这种倾向性影响的原因基础上,提出了基于遗传交叉运算的改进方法,针对小类别训练样本,利用交叉运算产生新的样本,从而补偿了不平衡样本数据所造成的影响,并减少了训练模型的误差,这对于某些需要重点关注的小类别精度的应用场合有非常重要的现实意义.基于 UCI 标准数据集的仿真结果表明,改进算法比标准支持向量机具有更好的分类准确率.在今后的研究中,将进一步分析数据集中每一类别的识别准确率,并将其与其他方法进行更加充分的比较分析.

参考文献:

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer2Verlag. 2000. 138 - 167.

[2] Chew Hong-gunn, Crisp D J, Bogner R E, et al. Target detection in radar imagery using support vector machine with training size biasing[A]. Sundararajan N, Proceeding of the sixth International Conference on Control, Automation, Robotics and Vision[C]. Singapore: Proceedings: CD-ROM. 2000.

[3] Chew Hong-gunn, Bogner Robert E, Lim Cheng-chow. Dual nu-support vector machine with error rate and training size biasing[A]. V John mathews. Proceedings of 26th IEEE ICASSP (international Conference on Acoustics, Speech, and Signal Processing)[C]. Salt Lake city, UT, USA: IEEE, 2001. 1269 - 1272.

[4] Scholkopf B, Smola A, Williamson R C, et al. New support vector algorithm[J]. Neural Computation, 2000, 12(5): 1207 - 1245.

[5] Lin Chi-fu, Wang Shang-de. Fuzzy support vector machines[J]. IEEE Transaction on Neural Networks, 2002, 13(2): 464 - 471.

[6] 范昕炜,杜树新,吴铁军. 可补偿类别差异的加权支持向量机算法[J]. 中国图像图形学报, 2003, 18A(9): 1037 - 1042.
Fan Xin-wei, Du Shu-jun, Wu Tie-jun. Weighted support vector

machine based classification algorithm for uneven class size problem[J]. Journal of Image and Graphics, 2003, 18A(9): 1037 - 1042. (in Chinese)

[7] 周明,孙树栋. 遗传算法原理及应用[M]. 北京:国防工业出版社,1999. 45 - 54.

[8] Murphy P M, Aha Irvine DW CA. University of California, Department of Information and Computer Science [EB/OL]. http://www.ics.uci.edu/～mlearn/MLRepository.html. 1994.

[9] J Ma, Y Zhao, S Ahalt: OSU SVM Classifier Matlab Toolbox (version 3.00) [EB/OL]. Ohio State University, Columbus. USA. http://www.ece.osu.edu/～maj/osu svm/. 2002.

[10] The Genetic Algorithm Toolbox for Matlab[EB/OL]. Department of Automatic Control and Systems Engineering of The University of Sheffield, UK. http://www.shef.ac.uk/acse/research/ecrg/getgat.html.

作者简介:

吴洪兴 男,1973 年生于黑龙江省双城市,哈尔滨工业大学自动化测试与控制系博士生,主要研究领域为模式识别和智能故障诊断理论. E-mail: wuhongxing@hit.edu.cn

彭 宇 男,1973 年生于西安,哈尔滨工业大学自动化测试与控制系副教授,主要研究领域为计算智能和智能故障诊断理论.