文章编号:1001 - 9081(2011)04 - 1114 - 03
doi:10.3724/SP.J.1087.2011.01114

基于信息熵的支持向量数据描述分类

何伟成,方景龙
(杭州电子科技大学 计算机学院,杭州 310018)
(heweicheng@hotmail.com)

摘　要:针对现有的支持向量数据描述(SVDD)在解决分类问题时通常存在盲目性和有偏性,在研究信息熵和SVDD分类理论的基础上,提出了改进两类分类问题的E-SVDD算法。首先对两类样本数据分别求出其熵值;然后根据熵值大小决定将哪类放在球内;最后结合两类样本容量以及各自的熵值所提供的分布信息,对SVDD算法中的C值重新进行定义。采用该算法对人工样本集和UCI数据集进行实验,实验结果验证了算法的可行性和有效性。

关键词:信息熵;分布特性;支持向量数据描述;分类
中图分类号: TP181　　**文献标志码**:A

Classification method for SVDD based on information entropy

HE Wei-cheng, FANG Jing-long
(*School of Computer Science, Hangzhou Dianzi University, Hangzhou Zhejiang* 310018, *China*)

Abstract: Most of Support Vector Data Description (SVDD) methods have blindness and bias issues when working on two-class problems. The authors proposed a new SVDD method based on information entropy. In this algorithm, firstly, the entropy values were resolved respectively of the two classes of samples. Secondly, according to the size of the value, one class was placed inside the ball. Finally, the penalty was given based on the information provided by the sizes of the two sample data and their entropy values. The efficiency of this algorithm was verified by using artificial data and UCI datasets for the data imbalanced classification problem. The experimental results on artificial data sets and UCI data sets show the feasibility and effectiveness of the proposed method.

Key words: information entropy; distribution character; Support Vector Data Description (SVDD); classification

支持向量数据描述(Support Vector Data Description, SVDD)是Tax等人[1]在支持向量机(Support Vector Machine, SVM)基础上提出的一种数据描述的算法,该算法具有完善的理论支撑和稳定优良的性能表现,在机器学习和模式识别领域得到了广泛的应用。最初的SVDD方法是针对单类别的分类提出的,为了提高分类器的鲁棒性,Tax等人[2]在训练集中引入负类样本,使SVDD扩展为适用于两类分类问题。Mu等人[3]进一步改进了SVDD,将两类分类方法扩展到多类别分类。Chen等人[4]通过调整受试工作特征(Receiver Operating Characteristic, ROC)面积优化SVDD分类精度。但是目前的SVDD都将正类放在球内,有一定的盲目性;对惩罚值的设置也仅仅考虑将样本的数量信息进行调整。本文在SVDD算法的基础上,以各类样本提供的信息熵为包围球指导,将熵值小的类放在球内,然后根据信息熵所反映的样本分布信息与样本数量信息统一起来进行样本惩罚值的调整。实验结果证明,本文提出的基于信息熵的支持向量数据描述(Entropy-Support Vector Data Description, E-SVDD)算法与Tax等人[1-2]提出的适用于两类分类的SVDD算法相比,性能和分类精度都有较大的提高。

1　两类分类的支持向量域描述

Tax等人[1]建立了SVDD算法,适用于解决极端的不平衡分类学习即单分类学习问题。而适用于两类分类的SVDD基本思想是:在经过核映射的高维空间里构造一个几乎包含所有正类样本的球体,而将负类样本尽量排除在球外。

设样本集$(\boldsymbol{x}_i, y_i) \in \mathbf{R}^d$,其中$\boldsymbol{x}_i \in \mathbf{R}^d, y_i \in \{-1, +1\}$为对应类标,$i = 1,2,\cdots,n$。适用于两类分类的SVDD算法将球内的正类样本和球外的负类样本分别设置惩罚值为C^+、C^-,最优化问题为:

$$\min L = R^2 + C^+ \sum_r \xi_r + C^- \sum_l \xi_l \tag{1}$$

约束条件为:

$$\| \boldsymbol{x}_r - \boldsymbol{a} \|^2 \leqslant R^2 + \xi_r, \| \boldsymbol{x}_l - \boldsymbol{a} \|^2 > R^2 - \xi_l;$$
$$\xi_r \geqslant 0, \xi_l \geqslant 0 \tag{2}$$

其中:$\boldsymbol{x}_r$为正类样本;$\boldsymbol{x}_l$为负类样本;$\boldsymbol{a}$为超球球心;R为超球半径;C^+、C^-为惩罚因子,比例关系为$\frac{C^+}{C^-} \propto \frac{n_- v_-}{n_+ v_+}$,$n_+$为正类样本大小,$n_-$为负类样本大小,$v_+$为预设正类拒识率,$v_-$为预设负类误识率;$\xi_i$、$\xi_l$为松弛变量。

上述优化问题的求解请参考文献[5]。本文选用Gaussian径向基函数(Radial Basis Function, RBF)核函数:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left[-\frac{\| \boldsymbol{x} - \boldsymbol{y} \|^2}{2\sigma^2} \right] \tag{3}$$

在SVDD中,C值只与每类样本的个数和预设每类样本的拒识率有关,这样的C值设置对分布差异很大的样本分类结果影响很大。

收稿日期:2010 - 09 - 29;**修回日期**:2011 - 01 - 25。
基金项目:国家自然科学基金资助项目(60874074);浙江省科技计划重点项目(2009C14032)。
作者简介:何伟成(1986 -),男,浙江金华人,硕士研究生,主要研究方向:模式识别、支持向量机;　方景龙(1964 -),男,江西景德镇人,研究员,主要研究方向:模式识别、支持向量机。

2　基于信息熵的 E-SVDD

SVDD 的目标是对样本建立最小包围超球,实现对某一类样本的最小体积描述,即对包围的这一类样本的体积最小化压缩——最小化类内体积,本质上等价于类内聚类性的最大化[6]。但是这样的 SVDD 没有考虑样本的分布特性,在许多实际问题中,每类样本点由于其所处区域的分布不同其重要性也不同,从聚类的角度看,信息熵小的类比信息熵大的类有更好的聚类性[7],更应包含在超球体内。

惩罚因子的作用是调节学习机器置信范围和经验风险的比例以使学习机器的推广能力最好。通常机器学习对正负样本分别采用不同惩罚参数,但目前惩罚参数设置只考虑样本数量信息,即样本数量较多的类别惩罚参数较小,样本数量较少的类别惩罚参数较大,而没有考虑样本分布特性对分类的影响。由文献[8]分析认为,信息熵能反映类的分布特性。从分类识别的角度看,熵值越小,表明类的方差越小,分类特性越好,该类也能提供较好的分类信息。为了防止其被误分,只要对其施加更大的惩罚系数即可。相反,熵值越大,给该类的惩罚系数越小。由上述分析可以得到不同类惩罚因子的比例关系 $\frac{C^+}{C^-} \propto \frac{H(-)}{H(+)}$。

2.1　信息熵

信息熵是对事物状态有序性的一种度量。某一事物状态不确定性的大小,与该事物可能出现的状态数目以及各状态出现的概率有关。熵值越大,说明系统中的数据越无序,系统越杂乱;反之,熵值越小,则说明系统中的数据越有序,系统越纯净。

定义 1　令 X 代表一个属性,其取值的集合为 Ψ,X 取值的概率分布为 $p(x)$,则属性 X 的信息熵[9]为:

$$H(X) = E[-\mathrm{lb}\, p(x)] = -\sum_{x\in\Psi} p(x)\mathrm{lb}\, p(x) \tag{4}$$

定义 2　令 $\hat{X} = \{X_1, X_2, \cdots, X_d\}$ 表示多个属性,它的概率函数为 $p(x_1, x_2, \cdots, x_d)$,则信息熵表示如下:

$$H(\hat{X}) = -\sum_{x_1\in\Psi_1}\cdots\sum_{x_d\in\Psi_d} p(x_1,\cdots,x_d)\ \mathrm{lb}\, p(x_1,\cdots,x_d) \tag{5}$$

如果属性之间相互独立,式(5)可以转化成式(6)。为了简化对信息熵的计算,在本文中均假设数据集中的属性间是相互独立的。

$$\begin{aligned} H(\hat{X}) &= -\sum_{x_1\in\Psi_1}\cdots\sum_{x_d\in\Psi_d}[(p(x_1)\cdots p(x_d))\cdot \\ &\quad \mathrm{lb}\,(p(x_1)\cdots p(x_d))] = \\ &\quad H(X_1) + H(X_2) + \cdots + H(X_d) \end{aligned} \tag{6}$$

2.2　C 值的确定

在前面分析的基础上,将样本数量信息与样本分布信息统一起来进行样本惩罚值调整。两类样本分别采用不同的惩罚因子 C^+, C^-,由前面可知它们的比例 $\frac{C^+}{C^-} \propto \frac{n_- v_-}{n_+ v_+}$。结合信息熵提供的信息,得到惩罚因子比例为 $\frac{C^+}{C^-} \propto \frac{H(-)n_- v_-}{H(+)n_+ v_+}$。

在 E-SVDD 中,样本的拒识率预先设置,同时考虑到正负两类样本的分布和数量的不同,为两类样本各定义一个 C 值:

$$C^+ = \frac{1}{H(+)n_+ v_+} \tag{7}$$

$$C^- = \frac{1}{H(-)n_- v_-} \tag{8}$$

结合以上分析,考虑分布特性,并结合样本容量差异,给出了 E-SVDD 算法。该算法分三步进行:

1)利用信息熵分别求出正类与负类的信息 $H(+)$、$H(-)$,将熵值小的类放在超球内;

2)利用两类样本容量差异以及信息熵所提供的信息来确定两类数据的惩罚因子比例;

3)建立模型求解。

3　实验结果与分析

3.1　实验设置

本文共采用了 7 个数据集,即 3 个人工数据集 a1、a2、a3 和 4 个 UCI 数据集 Iris、Glass、Delft、Balance,具体信息见表 1,其中 n_+ 是正类样本的个数,n_- 是负类样本的个数。

表 1　实验中使用的数据集

数据集	n_+	n_-	属性
a1	100	100	2
a2	50	100	2
a3	50	100	2
Iris	50	50	4
Glass	70	144	9
Imports	71	88	25
Balance	288	337	4

本文采用 ROC 曲线下面积(Area Under ROC Curve,AUC)[10]来评价分类器的性能。AUC 即 ROC 曲线下的面积,一个有效的分类器的 AUC 应该大于 0.5,AUC 值越大则对应分类器的性能越好。文中所采用 SVDD 算法来自 Tax 等人[1]编写的 Dd_tools 工具箱,E-SVDD 是自己编写的算法。所有实验均在 Matlab 7.8 平台上进行,均采用 5 次 5 倍交叉验证,同时考虑取样的随机性,实验结果取 5 次 5 倍交叉验证的均值。

3.2　实验结果分析

首先对采用的各数据集的各类求信息熵,计算得到正类的熵值 $H(+)$ 和负类的熵值 $H(-)$ 以及惩罚值 C^+、C^- 的比值,结果见表 2。

表 2　信息熵的计算结果

数据集	$H(+)$	$H(-)$
a1	6.80	7.75
a2	6.52	7.74
a3	7.38	5.07
Iris	14.76	15.63
Glass	40.66	47.67
Imports	75.75	74.82
Balance	8.70	8.87

为了验证将信息熵小的类放在球内分类效果更好,使用原 SVDD 算法分别将正类放在球内、负类放在球外与负类放在球内、正类放在球外,记为 SVDD(+)和 SVDD(-)。下面实验是在 7 个数据集下使用不同核参数时 AUC 值比较,核参数设置为 1~10,球内拒识率都设为 0.1,球外拒识率为 1,评价标准取 AUC 值。实验结果如表 3 所示。

由表 3 结果可得:7 个数据集的实验结果均表明将熵值小的类放在球内比熵值大的类放在球内得到分类精度更高。这与熵值小的类更适合放在球内的分析是一致的,也更满足 SVDD 的学习策略。在接下来的实验中均采用将熵值小的类放在球内的学习策略。

然后根据样本数量信息与样本分布信息确定惩罚 C 值,

在 SVDD 基础之上改进后形成本文的 E-SVDD 算法。下面的实验是 SVDD 和 E-SVDD 在 7 个数据集上使用不同核参数时的性能比较,核参数设置为 1 ~ 10,评价标准取 AUC 值。实验结果如表 4 所示。

从表 4 中可以看出,从整体 AUC 值方面分析,除下划线字体以外,所有数据集的实验结果均说明 E-SVDD 算法比 SVDD 算法更优,而且随着核参数的增加,E-SVDD 表现得比 SVDD 稳定,可见本文提出的算法性能比原 SVDD 更优、更稳定。同时,从数据集 a1 的结果分析,当核参数为 2 时,E-SVDD 算法取得的 AUC 值比原 SVDD 小 0.003,说明本文的算法不仅与所选的数据集相关,也与所选的核参数相关。从整体实验结果也说明,两个分类器对核参数是敏感的,分类效果受核参数的影响较大。因此,在实际应用时核参数的选择很关键。

表 3 SVDD(+) 与 SVDD(-) 的 AUC 值实验结果

数据集	方法	核参数的取值									
		1	2	3	4	5	6	7	8	9	10
a1	SVDD(+)	0.869	0.879	0.871	0.885	0.899	0.902	0.900	0.911	0.905	0.916
	SVDD(-)	0.803	0.802	0.801	0.774	0.776	0.806	0.847	0.855	0.857	0.886
a2	SVDD(+)	0.917	0.918	0.914	0.916	0.907	0.913	0.904	0.907	0.916	0.913
	SVDD(-)	0.765	0.721	0.749	0.761	0.788	0.789	0.825	0.845	0.858	0.876
a3	SVDD(+)	0.526	0.541	0.593	0.614	0.682	0.719	0.759	0.802	0.812	0.778
	SVDD(-)	0.930	0.964	0.963	0.942	0.937	0.935	0.936	0.937	0.938	0.937
Iris	SVDD(+)	0.984	0.990	0.976	0.980	0.978	0.978	0.978	0.978	0.976	0.974
	SVDD(-)	0.892	0.926	0.938	0.950	0.960	0.962	0.960	0.960	0.960	0.960
Glass	SVDD(+)	0.815	0.774	0.766	0.759	0.751	0.744	0.736	0.754	0.764	0.772
	SVDD(-)	0.352	0.385	0.439	0.446	0.503	0.513	0.529	0.522	0.519	0.509
Imports	SVDD(+)	0.616	0.591	0.584	0.584	0.614	0.641	0.676	0.680	0.701	0.717
	SVDD(-)	0.680	0.687	0.713	0.762	0.773	0.780	0.767	0.747	0.750	0.745
Balance	SVDD(+)	0.891	0.942	0.956	0.957	0.954	0.955	0.939	0.948	0.938	0.947
	SVDD(-)	0.839	0.933	0.952	0.939	0.946	0.939	0.935	0.921	0.931	0.924

表 4 在 7 个数据集中使用不同核参数时 SVDD 与 E-SVDD 的 AUC 值比较

数据集	方法	核参数的取值									
		1	2	3	4	5	6	7	8	9	10
a1	SVDD	0.869	0.879	0.871	0.885	0.899	0.902	0.900	0.911	0.905	0.916
	E-SVDD	0.869	0.880	0.899	0.903	0.915	0.921	0.925	0.934	0.916	0.932
a2	SVDD	0.917	0.918	0.914	0.916	0.907	0.913	0.904	0.907	0.916	0.913
	E-SVDD	0.917	0.915	0.919	0.933	0.925	0.933	0.938	0.925	0.938	0.942
a3	SVDD	0.930	0.964	0.963	0.942	0.937	0.935	0.936	0.937	0.938	0.937
	E-SVDD	0.930	0.965	0.968	0.974	0.973	0.965	0.969	0.959	0.957	0.955
Iris	SVDD	0.984	0.990	0.976	0.980	0.978	0.978	0.978	0.978	0.976	0.974
	E-SVDD	0.984	0.990	0.984	0.980	0.978	0.978	0.978	0.978	0.978	0.978
Glass	SVDD	0.815	0.774	0.766	0.759	0.751	0.744	0.736	0.754	0.764	0.772
	E-SVDD	0.848	0.813	0.829	0.793	0.790	0.774	0.771	0.797	0.778	0.796
Imports	SVDD	0.680	0.687	0.713	0.762	0.773	0.780	0.767	0.747	0.750	0.745
	E-SVDD	0.680	0.687	0.713	0.764	0.777	0.799	0.782	0.817	0.804	0.770
Balance	SVDD	0.891	0.942	0.956	0.957	0.931	0.955	0.939	0.948	0.938	0.947
	E-SVDD	0.899	0.977	0.982	0.984	0.984	0.983	0.983	0.984	0.982	0.983

4 结语

本文针对现有的 SVDD 算法缺陷,提出了改进的 E-SVDD 算法。通过和原有 SVDD 算法进行实验比较,该算法在各方面均有所提高。在改进的算法中所提出的调整方法没有修改原始样本信息,提供了更有效、更科学的学习策略,使算法的鲁棒性更好。未来需要进一步改进算法的参数选取,优化运算时间,以应用于更大规模的数据处理。

参考文献:

[1] TAX D M J, DUIN R P W. Support vector domain description [J]. Pattern Recognition Letters, 1999, 20(11/12/13): 1191 - 1199.

[2] TAX D M J. One-class classification [D]. Delft, Netherlands: Delft University of Technology, 2001.

[3] MU T, NANDI A K. Multiclass classification based on extended support vector data description [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(5): 1206 - 1216.

[4] CHEN B, LI B, PAN Z S. SVDD regularized with area under the ROC [C] // Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human. Seoul: [s. n.], 2009: 364 - 368.

[5] TAX D M J, DUIN R P W. Support vector data description [J]. Machine Learning, 2004, 54(1): 45 - 66.

[6] 文传军, 詹永照, 陈长军. 最大间隔最小体积球形支持向量机[J]. 控制与决策, 2010, 25(1): 79 - 83.

[7] 熊家军, 李庆华. 信息熵理论与入侵检测聚类问题研究[J]. 小型微型计算机系统, 2005, 26(7): 1163 - 1166.

[8] 孙即祥, 姚伟, 滕书华. 模式识别[M]. 北京: 国防工业出版社, 2009.

[9] BABAIE T, LUCAS C. Variable selection using information entropy in time series prediction [C] // Proceedings of the Seventh International Conference on Computer Science and Information Technologies. Yerevan, Armenia: [s. n.], 2009: 118 - 121.

[10] TAX D M J, JUSZCZAK P. Kernel whitening for one-class classification [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2003, 17(3): 333 - 347.