文章编号:1000-5013(2013)05-0525-04 **doi**:10.11830/ISSN.1000-5013.2013.05.0525

一种基于数据分布的 SVM 核选择方法

郭金玲[1],王文剑[2,3]

(1. 山西大学 商务学院,山西 太原 030031;
2. 山西大学 计算机与信息技术学院,山西 太原 030006;
3. 山西大学 计算智能与中文信息处理教育部重点实验室,山西 太原 030006)

摘要: 针对目前支撑向量机(SVM)核函数的选择没有统一规则的现状,提出一种结合数据分布特征进行 SVM 核选择的方法.首先,采用多维尺度(MDS)分析方法对高维数据集合理降维,提出判断数据集是否呈圆球分布的算法;然后,在得到数据集分布特征的基础上进行 SVM 核选择,以达到结合数据分布特征合理选择 SVM 核函数的目的.实验结果表明:呈圆球分布的数据集采用球面坐标核进行分类,识别率达到 100%,训练时间最短,优于采用高斯核 SVM 及多项式核 SVM 的分类效果.

关键词: 支撑向量机;核函数;核选择;数据分布;多维尺度
中图分类号: TP 301 **文献标志码**: A

支撑向量机(support vector machine,SVM)最早由 Vapnik 提出[1],是一种基于核的机器学习方法,主要用于解决数据的分类与回归问题.不同的核函数对 SVM 的泛化能力有重要的影响,目前常用的核函数有高斯核[2]和多项式核[3].所以要构造 SVM,首先要进行 SVM 核选择[4-12].核函数的选择没有统一的规则,凭经验选取,因此核选择是使用 SVM 算法首要的步骤,如何有效进行核选择是 SVM 的一个重要研究内容.分类问题中,SVM 核函数选择的方法主要分为基于数据依赖的方法和基于数据独立的方法.数据依赖的核函数选择方法一般基于先验数据,SVM 训练之前对核及参数进行优化处理.如周伟达等[13]提出的极小化 R^2/Δ^2 的核选择方法;Amari 等[14]提出两步迭代法等.数据独立的方法主要利用有关问题的先验信息进行 SVM 核函数选择,代表性的方法有留一交叉校验法(leave-one-out cross validation)[15]和基于 VC 维界的估计方法[16]等.数据依赖的方法一般计算代价小,可解决很多领域的任何问题,具有通用性,但容易产生过拟合问题,泛化能力较差;而数据独立的方法计算代价太大,不实用,一般只作为参考.目前大多数核选择方法都不考虑数据的分布特征,没有充分利用隐含在数据中的信息.如果已知数据的分布特征或可得到其数据分布特征的近似,再进行 SVM 核选择,可以在很大程度上提高 SVM 的泛化能力.据此,本文提出一种结合数据集几何分布特征进行核选择的方法.

1 数据预处理

由于实际问题中大多数据集都是高维数据集,为了判断其近似的几何形状,同时解决高维数据带来的维数灾难问题,首先对数据集进行降维.多维尺度(multidimensional scaling,MDS)分析方法[17]是一种把原来多个变量划为少数几个综合指标的降维处理方法,是一种线性降维方法.设原始高维数据集 $X=\{x_1,x_2,\cdots,x_l\}$,维数为 k,MDS 算法有如下 5 个具体步骤.

步骤 1 根据不相似度 $\delta_{i,j}^2$ 得到不相似度矩阵 $\boldsymbol{A}$,即 $\boldsymbol{A}=[-\frac{1}{2}\delta_{i,j}^2]$,$\delta_{i,j}^2=(x_i-x_j)^{\mathrm{T}}(x_i-x_j)$.其中,$i=1,2,3,\cdots,l$; $j=1,2,3,\cdots,l$.

收稿日期: 2013-03-01

通信作者: 郭金玲(1982-),女,讲师,主要从事机器学习与数据挖掘的研究.E-mail:tygjl@163.com.

基金项目: 国家自然科学基金资助项目(61273291);山西省高等学校科技研究开发项目(20121131);山西大学商务学院科研基金资助项目(2012013)

步骤 2 计算中心化矩阵 $\boldsymbol{A}$,$\boldsymbol{B}=\boldsymbol{HAH}$,$\boldsymbol{H}=\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{l}\boldsymbol{l}^{\mathrm{T}}$. 其中 $\boldsymbol{H}$ 是中心矩阵,$\boldsymbol{I}_n$ 是单位矩阵.

步骤 3 计算矩阵 $\boldsymbol{B}$ 的特征值及特征向量,即$[\boldsymbol{e},\lambda]=\mathrm{eig}(\boldsymbol{B})$. 将特征值 λ_i 排序,即 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_k$,其中 λ_i 对应的特征向量是 $\boldsymbol{e}_i$. 由此可得到尺度集$\{S_i\mid S_i=\boldsymbol{e}'_iX\}(i=1,2,\cdots,k)$.

步骤 4 数据集降维. 元素 S_i 的贡献率为 $P_i=\lambda_i/\sum_{n=1}^{k}\lambda_n,(i=1,2,\cdots,k)$,前 m 个元素的累计贡献率为 $P=\sum_{n=1}^{i}\lambda_n/\sum_{n=1}^{k}\lambda_n,(i=1,2,\cdots,m)$.

一般地,降维后的输出数据集取累计贡献率达 90%～95%的特征值,$\lambda_1,\lambda_2,\lambda_3,\cdots,\lambda_m$ 所对应的尺度集中的 $m(m\Leftarrow k)$个元素.

步骤 5 输出降维后数据集 X'. $X'=\{z_1,z_2,\cdots,z_l\}$,维数为 m.

由以上算法分析得到降维后的 m 维数据集 $X'=\{z_1,z_2,\cdots,z_l\}_{m\times l}$为

$$X'=\{z_1,z_2,\cdots,z_l\}=\mathrm{diag}(\lambda_1^{1/2},\cdots,\lambda_m^{1/2})\boldsymbol{U}^{\mathrm{T}},\qquad \boldsymbol{U}=\{e_1,e_2,\cdots,e_m\}.$$

文献[17]给出了 MDS 算法保距性定理,即低维数据集 X'满足

$$X'=\arg\min_{L}\|\boldsymbol{L}^{\mathrm{T}}-\boldsymbol{H}\|.$$

其中:$\boldsymbol{L}$ 为任意的 $m\times l$ 矩阵;$\boldsymbol{H}=X^{\mathrm{T}}X$. 定理的具体证明过程参阅文献[17]. 由该定理可知:利用 MDS 算法对数据进行降维的过程中,最大限度地保持了数据点在原始空间的距离,从而最大程度地保持了原始数据集的几何分布特征.

2 核选择方法

设实验数据集包含两类样本,分别是 A 类样本和 B 类样本. 数据集呈圆球分布,是指一类样本在圆球内,另外一类样本在圆球外. 基于数据分布的 SVM 核选择方法有如下 3 个具体步骤.

步骤 1 对高维数据集,采用 MDS 方法降维处理成三维数据集.

步骤 2 判定实验数据集是否呈圆球分布,如图 1 所示. 设 A 类样本的重心为 O,d_{A} 和 d_{B} 为存放 A 类和 B 类样本各点到 O 的距离,其最大值记为 $d_{\mathrm{A,max}}$和 $d_{\mathrm{B,max}}$,最小值记为 $d_{\mathrm{A,min}}$和 $d_{\mathrm{B,min}}$.

步骤 3 结合样本集的分布选择相应的核函数,样本集呈圆球分布,则 SVM 选择球面坐标核;反之,选择常用的高斯核或多项式核.

该方法结合实验数据集的几何分布进行核选择,可以降低计算代价,能够较好体现数据的分布特征,直观性较强.

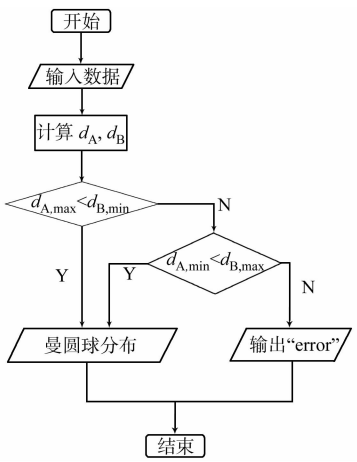


图 1 圆球分布判定流程图

Fig. 1 Flowchart of deciding sphericity distribution

3 实验结果分析

实验数据采用人工构造的圆球数据集 $D1$,随机生成 100 个样本,其分布如图 2 所示. 图 2 中:数据集 $D1$ 包含两类样本,A 类样本满足条件 $x^2+y^2+z^2\leqslant1$,用圆圈表示;B 类样本满足条件 $x^2+y^2+z^2>1$,用星号表示. 采用核选择方法对数据集 $D1$ 进行检测,可得 $D1$ 呈圆球分布.

采用球面坐标核进行分类实验,不涉及参数选取. 实验中,从两类样本中随机选取 40 个点作为训练样本,剩余的数据点作为检验样本,进行了 12 次数值实验,取平均结果作为最后结果,如图 3(a)所示.

采用高斯核进行分类实验,参数 σ 分别取 0.1,0.8,1,2,10 进行实验,其中参数 $\sigma=0.8$ 时的分类效果较好. 取参

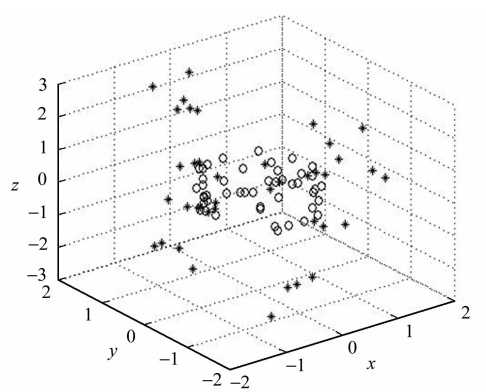
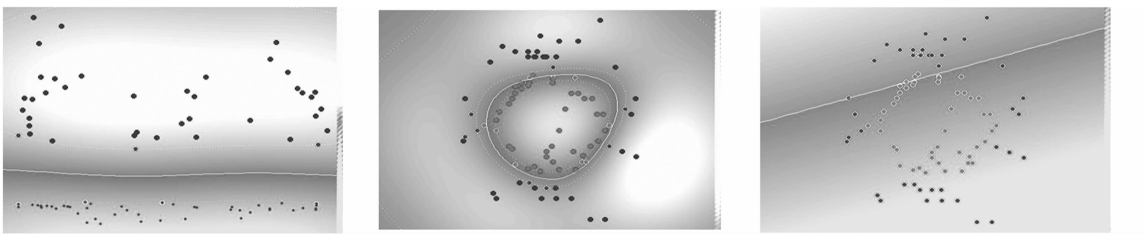


图 2 数据集 $D1$ 的分布情况

Fig. 2 Distribution of dataset $D1$

数 $\sigma=0.8$，从两类样本中随机选取 40 个点作为训练样本，剩余的数据点作为检验样本，进行了 12 次数值实验，取平均结果作为最后结果，如图 3(b)所示.

采用多项式核进行分类实验，参数 d 分别取 1，2，8，10，15 进行了实验，其中 $d=2$ 时的分类效果较好. 取 $d=2$，从两类样本中随机选取 40 个点作为训练样本，剩余的数据点作为检验样本，进行了 12 次数值实验，取平均结果作为最后结果，如图 3(c)所示.



(a) 球面坐标核 SVM　　(b) 高斯核 SVM　　(c) 多项式核 SVM

图 3　不同分类方法对数据集 $D1$ 的分类

Fig. 3　Classification results on $D1$ by different ways

采用不同方法对 $D1$ 进行分类，结果如表 1 所示. 表 1 中：N 为样本总数；n 为错分样本数；η 为识别率；t 为训练时间. 从实验结果可以看出：对于呈圆球分布的数据集 $D1$，采用球面坐标核进行分类实验，识别率达到 100%，训练时间最短，且分类效果优于采用高斯核 SVM 及多项式核 SVM 的效果. 进一步分析，由文献[4]可知，运用球面坐标核进行空间直角坐标到球面坐标的同维映射，映射 ϕ 满足以下关系

表 1　不同分类方法对数据集 $D1$ 的分类比较

Tab. 1　Comparisons of classification results on $D1$ by different ways

分类方法	N	n	η/%	t/s
球面坐标核 SVM	100	0	100	1.9
高斯核 SVM	100	0	100	12.0
多项式核 SVM	100	30	70	20.1

$$(x_1,x_2,x_3)^{\mathrm{T}} \xrightarrow{\phi} \begin{pmatrix}\psi\\ \theta\\ r\end{pmatrix} = \begin{pmatrix}\arccos(x_3/\sqrt{x_1^2+x_2^2+x_3^2})\\ \operatorname{arctg}(x_2/x_1)\\ \sqrt{x_1^2+x_2^2+x_3^2}\end{pmatrix}.$$

经过计算，显然满足条件 $x^2+y^2+z^2\leqslant 1$ 的 A 类样本经过映射 ϕ，在特征空间均分布在平面 $r=1$ 下方；满足条件 $x^2+y^2+z^2>1$ 的 B 类样本经过映射 ϕ，在特征空间均分布在平面 $r=1$ 上方，即分界面趋于平面.

该分析具有一般性，按照文中提出的算法检测得到的呈圆球分布，其两类数据集均有明显的分界面，且该分界面在特征空间将转化为一个平面. 因此，呈圆球分布的数据集采用球面坐标核进行分类实验，分类效果较好.

4　结束语

分析了数据集呈圆球分布时 SVM 核函数选择的方法，实验中分别采用 3 种不同核的 SVM 对数据集进行分类. 结果表明：数据集呈圆球分布时，采用球面坐标核 SVM 训练速度快且分类效果好，证明方法的正确性、有效性，提高了 SVM 的泛化能力.

参考文献：

[1] VAPNIK V. The nature of statitiscal learning theory[M]. New York: Spring Verlag Press, 1995: 4-15.

[2] WANG Wen-jian, XU Zong-ben, LU Wei-zhen, et al. Determination of the spread parameter in the Gaussian kernel for classification and regression[J]. Neurocomputing, 2003, 55(3/4): 643-663.

[3] 孙建涛，郭崇慧，陆玉昌，等. 多项式核支持向量机文本分类器泛化性能分析[J]. 计算机研究与发展，2004，41(8)：1321-1326.

[4] 张莉，周伟达，焦李成. 一类新的支撑矢量机核[J]. 软件学报，2002，13(4)：713-718.

[5] WANG Jin-jun, YANG Jian-chao, YU Kai, et al. Locality constrained linear coding for image classification[J].

CVPR,2010,15(3):456-470.
[6] WANG Xiao-ming,CHUNG Fu-lai,WANG Shi-tong. Theoretical analysis for solution of support vector data description[J]. Neural Networks,2011,24(4):360-369.
[7] ZAFEIRIOU S,TEFAS A,PITAS I. Minimum class variance support vector machines[J]. IEEE Transanctions on Image Processing,2007,16(10):2551-2564.
[8] ZHOU Xi,CUI Na,LI Zhen,et al. Hierarchical gaussianization for image classification[J]. ICCV,2009,18(3):79-90.
[9] HUANG Kai-zhu,YANG Hai-qing,KING I,et al. Maxi-min margin machine: Learning large margin classifiers locally and globally[J]. IEEE Transanctions on Neural Networks,2008,19(2):260-272.
[10] YU Kai,ZHANG Tong,GONG Yi-hong. Nonlinear learning using local coordinate coding[J]. NIPS,2009,26(8):342-356.
[11] CHOI Y S. Least squares one-class support vector machine[J]. Pattern Recognition Letters,2009,30(13):1236-1240.
[12] GAO Sheng-hua,TSANG I W H,CHIA L T,et al. Local features are not lonely laplacian sparse coding for image classification[J]. CVPR,2010,18(6):126-138.
[13] 周伟达,张莉,焦李成. 一种改进的推广能力度量标准[J]. 计算机学报,2003,26(5):598-604.
[14] WU Si,AMARI S I. Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers[J]. Neural Processing Letters,2002,15(1):59-67.
[15] CHAPELLE O,VAPNIK V. Model selection for support vector machines[C]// Advances in Neural Information Processing Systems 12. Cambridge:MIT Press,2001:120-155.
[16] VAPNIK V. The nature of statitiscal learning theory[M]. New York:Spring-Verlag Press,1995:88-125.
[17] COX T,COX M. Multidimensional Scaling[M]. London:Chapman & Hall,1994:145-150.

A SVM Kernel Selection Approach Based on the Characteristics of Data Distribution

GUO Jin-ling[1], WANG Wen-jian[2,3]

(1. Business College of Shanxi University, Taiyuan 030031, China;
2. School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China;
3. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)

Abstract: The kernel selection has no unified rules for support vector machine (SVM). Based on the characteristics of dataset distribution, a new way to select the kernel function was presented. First dimension reduction of the high dimensional dataset was processed with multidimensional scaling (MDS) method. Then an algorithm was put forward, it was judged whether dataset is sphericity distribution. On the basis of determining sphericity distribution, how to select the kernel function was discussed, to achieve the purpose of selecting SVM kernel function with data distribution characteristics. The experimental results illustrate that the classification recognition rate of sphericity datasets reaches 100% with sphere kernel and the training time is the shortest. The classification effect is better than that of using gaussian kernel SVM and polynomial kernel SVM.

Keywords: support vector machine; kernel function; kernel selection; data distribution; multidimensional scaling

(**责任编辑**: 黄晓楠 **英文审校**: 吴逢铁)