基于邻域粗糙集的支持向量机分类方法研究

韩 虎[1,2] 党建武[2] 任恩恩[2]
(兰州交通大学数理与软件学院 兰州 730070)[1] (兰州交通大学电子与信息工程学院 兰州 730070)[2]

摘 要 针对支持向量机方法对高维大规模数据无法直接处理和对异常样本敏感的问题,提出了一种基于邻域粗糙集模型的改进支持向量机。该算法从两个方面对训练样本集进行预处理:一方面利用邻域粗糙集模型中对象邻域的上、下近似,寻找两种类别的交界部分,从而减小问题规模;然后通过对交界部分样本进行混淆度分析,剔除那些混杂在另一类样本中的异常样本或噪声数据。另一方面利用属性重要性度量对样本集进行属性约简与属性加权处理。基于合成数据集与标准数据集的有关实验证实了该算法的有效性。

关键词 支持向量机,邻域粗糙集,预处理,属性约简

Research of Support Vector Classifier Based on Neighborhood Rough Set

HAN Hu[1,2] DANG Jian-wu[2] REN En-en[2]
(School of Mathematics, Physics and Software Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)[1]
(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)[2]

Abstract Support vector machine can not directly deal with high dimension and large scale training set and it is sensitive to abnormal samples, an improved support vector classifier based on neighborhood rough set was proposed. In the paper, data preprocessing was done on training set from two different sides. On the one hand, neighborhood rough set was used to find these samples in boundary and obtain a reduced training set, at the same time, those abnormal samples which not only lead to over-learning but also decrease the generalization ability were deleted. On the other hand, attribute reduction was done and feature weight was imported based on attribute significance because different feature effects differently on classification. At last several comparative experiments using synthetic and real life data set show the performance and the effectivity of the method.

Keywords Support vector machine, Neighborhood rough set, Preprocess, Attribute reduction

支持向量机(Support Vector Machine,简称SVM)是一种基于统计学习理论的、专门研究有限样本预测的机器学习方法,建立在结构风险最小化基础上,与传统的学习方法相比具有较好的学习性能和泛化能力[1]。但是由于它存在高维大规模数据无法直接处理和对异常样本敏感的问题,使得支持向量机在很多情况下无法直接使用。

如何将支持向量机适用于大规模样本,一直是该领域的一个研究热点,其中一种重要的解决思路就是对可能是支持向量的样本进行预选取策略[2]。因为在支持向量机方法中,最大间隔分类超平面是由支持向量决定的,而支持向量恰恰位于两类别样本的交界部分,所以对于大规模训练集,可以通过预先找到两类别样本的交界部分来缩小样本集规模[3,4]。同时,由于支持向量机对噪声样本敏感,当两类训练样本集存在混淆或有噪声存在时,由交界部分样本训练得到的支持向量机分类面会变得过分复杂而使其泛化能力降低[5]。因此,对交界部分训练样本进行混淆度分析,剔除那些异常样本或噪声数据。

支持向量机通过一类核函数将输入空间非线性映射到特征空间,以$K(x,y)$代替$\phi(x)\cdot\phi(y)$,克服了维数灾难,使支持向量机方法在很多高维问题中得到了很好的应用。但是对于样本集中存在的冗余属性,它们不仅会影响支持向量机的推广能力,还会使支持向量机结构变得复杂,因此有必要对训练样本集进行属性约简,消除样本集中大量的冗余信息和冲突对象,从而提高支持向量机的推广能力。

本文一方面利用邻域粗糙集模型[6]中对象邻域的上、下近似,寻找两种类别的交界部分,从而减小问题规模;然后通过对交界部分样本进行混淆度分析,剔除那些混杂在另一类样本中的异常样本或噪声数据。另一方面利用属性重要性度量对样本集进行属性约简与属性加权处理,获得性能更优的改进支持向量机。

1 支持向量机原理

对于给定的一组样本集$\{x_i,y_i\}$,$i=1,2,\cdots,l$,这里$y_i=1$或-1,SVM依据结构风险最小化原则,将其学习过程转化为如下的优化问题:

$$\mathrm{Min}\ \frac{1}{2}\|w\|^2+C\sum_1^l\xi_i \tag{1}$$

到稿日期:2009-03-19 返修日期:2009-07-16 本文受甘肃省自然科学基金(2008GS02625),甘肃省教育厅科研基金(0804-01)资助。
韩 虎(1977—),男,博士生,主要研究方向为机器学习、数据挖掘,E-mail:hanhu-lzjtu@163.com;党建武 男,教授,博士生导师,主要研究方向为神经网络、智能计算;任恩恩 男,教授,博士生导师,主要研究方向为网格理论、复杂计算。

满足约束条件：

$$y_i(w \cdot z_i + b) \geqslant 1 - \zeta_i \quad (2)$$

$$\zeta_i \geqslant 0, i = 1, \cdots, l$$

其中，训练样本 x_i 被函数 $z_i = \phi(x_i)$ 映射到高维特征空间，$w \in R^N$ 是超平面的系数向量，$b \in R$ 为阈值，ξ_i 为松弛变量，$C > 0$ 是对错分样本的惩罚因子。

采用拉格朗日乘子法把上述最优分类面问题转换为其对偶问题，即寻找最大化目标函数：

$$\max(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \quad (3)$$

满足约束条件：

$$\sum_{i=1}^{l} \alpha_i y_i = 0, 0 \leqslant \alpha_i \leqslant C, i = 1, 2, \cdots, l \quad (4)$$

于是相应的分类决策函数为：

$$f(x) = \mathrm{sign}(\sum_{i=1}^{m} \alpha_i^* y_i K(x_i \cdot x) + b^*) \quad (5)$$

其中，α_i^* 为对应 $\alpha_i \neq 0$ 的向量，称为支持向量；$m(m < l)$ 为支持向量的数目；b^* 为与 α_i^* 对应的阈值；$K(x_i, x) = \phi(x_i) \cdot \phi(x)$ 为满足 Mercer[7] 条件的核函数。常用的 3 种核函数如表 1 所列。

表 1　常用核函数

Kernel function	Expression
Linear kernel	$x_i^T x$
Polynomial kernel	$(1 + x_i^T x)^d$
RBF kernel	$\exp(-\|x - x_i\|^2 / \sigma^2)$

2　邻域粗糙集

邻域模型概念是由 T. Y. Lin 于 1988 年提出来的[8]，该模型通过空间点的邻域来粒化论域空间。将邻域理解为基本信息粒子，用来描述空间中的其它概念。邻域粗糙集模型是胡清华等人利用邻域模型对经典粗糙集理论的一种拓展模型[6]，该模型以实数空间中的每一个点形成一个 δ 邻域，δ 邻域族构成了描述空间中任一概念的基本信息粒子。

对于信息系统 $IS = \langle U, A, V, f \rangle$，其中 $U = \{x_1, x_2, \cdots, x_n\}$ 表示非空有限集合，称为论域。A 是属性集合，V 是值域，$f: U \times A \to V$ 是一个信息函数，表示样本与其属性取值的对应映射关系。如果 $A = C \cup D$，其中 C 表示条件属性集，D 表示决策属性集，且要求 $C \cap D = \emptyset$，则 $IS = \langle U, A, V, f \rangle$ 称为一个决策表。对于 $x_i \in U$，定义 x_i 的邻域为：

$$\delta_B(x_i) = \{x_j \mid x_j \in U, \Delta_B(x_i, x_j) \leqslant \delta\}$$

其中，Δ 是一个距离函数。对于 $\forall x_1, x_2, x_3 \in U$，$\Delta$ 满足如下关系：

1) $\Delta(x_1, x_2) \geqslant 0$，$\Delta(x_1, x_2) = 0$，当且仅当 $x_1 = x_2$；

2) $\Delta(x_1, x_2) = \Delta(x_2, x_1)$；

3) $\Delta(x_1, x_3) \leqslant \Delta(x_1, x_2) + \Delta(x_2, x_3)$。

对于 N 个属性的样本集，距离常用 P 范数表示为：

$$\Delta_P(x_1, x_2) = (\sum_{i=1}^{N} | f(x_1, a_i) - f(x_2, a_i) |^P)^{1/P}$$

其中，$f(x, a_i)$ 为样本 x 在属性 a_i 上的取值。$\Delta_P(x_1, x_2)$ 定义是对于数值型属性集而言的，但邻域模型很容易将距离计算扩展到含有符号和数值型的数据上来，对于符号型属性 a_i，可以定义：

1) $| f(x_1, a_i) - f(x_2, a_i) | = 0$，若 x_1, x_2 在 a_i 上取值相同；

2) $| f(x_1, a_i) - f(x_2, a_i) | = 1$，若 x_1, x_2 在 a_i 上取值不同。

从而邻域粗糙集的下近似与上近似分别定义为：

$$\underline{NX} = \{x_i \mid \delta(x_i) \subseteq X, x_i \in U\}$$

$$\overline{NX} = \{x_i \mid \delta(x_i) \cap X \neq \emptyset, x_i \in U\}$$

则对应 X 的近似边界为 $BN(X) = \overline{NX} - \underline{NX}$。

对于一个邻域决策系统 $NDT = \langle U, C \cup D, V, f \rangle$，$D$ 将 U 划分为 N 个等价类：$X_1, X_2, \cdots, X_N$，$\forall B \subseteq C$，定义决策 D 关于 B 的下近似、上近似及决策边界分别为：

$$\underline{N_B D} = \bigcup_{i=1}^{N} \underline{N_B X_i}, \overline{N_B D} = \bigcup_{i=1}^{N} \overline{N_B X_i}, BN(D) = \overline{N_B D} - \underline{N_B D}$$

决策 D 的下近似，亦称为决策正域，记为 $\mathrm{POS}_B(D)$。正域的大小反映了分类问题在给定属性空间中的可分离程度，正域越大，表明各类的重叠区域，即边界越少。因此定义决策属性 D 对条件属性 B 的依赖性为：

$$\gamma_B(D) = \frac{| \mathrm{POS}_B(D) |}{| U |} \quad (6)$$

3　数据预处理

提高支持向量机效率，改善其推广能力的一种有效途径，就是对样本数据集进行预处理。把基于粗糙集理论的属性约简作为支持向量机的前端数据预处理器的研究已出现不少[9-11]，但都是从属性约简的角度进行处理。本文将同时从属性约简和样本选取两方面进行样本集的预处理。

3.1　属性约简

对于一个给定的信息系统，往往存在大量的冗余信息和冲突对象，它们不仅会影响支持向量机的推广能力，还会使支持向量机结构变得复杂。同时不同属性对支持向量机的贡献往往不同，因此有必要对训练样本集进行属性约简和加权处理。本文采用文献[12]描述的快速约简算法进行属性的约简操作，并利用条件属性对决策属性的依赖性度量，对各属性进行加权处理，具体步骤如下。

输入：决策表 $\langle U, C, D, V, f \rangle$

输出：约简 red

步骤 1　初始化 $red = \emptyset$，初始化待检验样本集 $sam\_chk = U$

步骤 2

```
flag=1;
while sam_chk≠∅
  for each k_i∈(C−red)
    DT_i=⟨U,red∪k_i,D,V,f⟩;
    初始化 POS_i=∅;
    for each a_j∈smp_chk
        计算 a_j 在 DT_i 下的邻域 δ(a_j);
        if δ(a_j)各样本决策属性 D 取值相同
        POS_i=POS_i∪a_j
        end if
    end for
    if flag=1
      γ_i=POS_i/smp_chk
  end for
  flag=0
  找出最大的 POS_i 和对应的 k_i;
  if POS_i≠∅
    red=red∪k_i
```

smp_chk=*smp_chk*−POS;

else

退出 while 循环

end if

end while

步骤 3　return *red*

步骤 4　对约简后的训练样本集进行属性加权，分别乘以相应的属性重要度 γ_i

3.2　样本选取

在邻域粗糙集模型中，边界部分样本常常表现为两类别样本的交界部分。图 1 给出一个二维空间的两类分类问题来对近似边界进行说明。

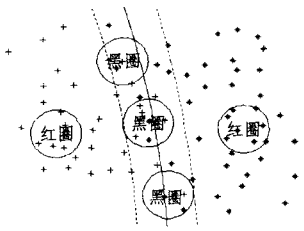

图 1　邻域粗糙集

以“+”标识的样本为第一类样本，以“*”标识的为第二类样本。我们看出红圈代表的圆形邻域内的样本都来自于同一类，因此它们属于下近似集。黑圈代表的圆形邻域内既有第一类的样本，也有第二类的样本，因此它们是边界部分样本。同时可以看到，几乎所有的支持向量都位于边界部分。因此，对于大规模样本，在训练之前利用邻域粗糙集模型找到边界部分样本，以边界部分样本为训练集，进行支持向量机训练，可以有效提高支持向量机的训练速度。

同时，边界部分样本也是异常样本和含噪声样本存在的地方。支持向量机方法对于这些样本都十分敏感，它们会使支持向量机分类超平面过度复杂，降低了分类器的分类效率和泛化性能。为了说明这一点，以图 2 和图 3 为例进行解释。

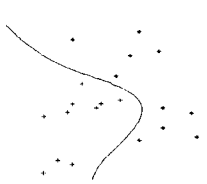

图 2　不存在噪声的分类结果

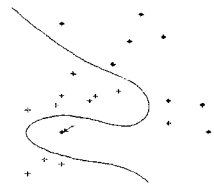

图 3　存在噪声的分类结果

图 2 所示的是对离心式水泵分别在正常状态和叶轮磨损状态下采集的一组压力样本的分类情况[13]。可以看出由它构造的分类超平面结构要比图 3 中所示的简单，这正是由于图 3 所示的是对一组故障样本在添加噪声的情况下所得的分类超平面，对于这类样本（图中箭头所指样本）应该在训练之前剔除。本文采用文献[14]定义的邻域匹配算子对得到的交界部分样本进行评估，具体定义如下：

$$\text{Neighbors\_Match}(x',k)=\frac{|\{x|\text{label}(x)=\text{label}(x'),x\in k\text{NN}(x')\}|}{k}$$

其中，$k\text{NN}(x')$ 是 x' 的 k 阶最近邻集合，Neighbors_Match (x',k) 的值越小，说明 x' 与其最近邻样本点的分布越不一致，它就越有可能是异常样本。因此在分析交界部分样本时，可以设定一个阈值 ε，当 Neighbors_Match$(x',k)<\varepsilon$ 时，对 x' 进行剔除，从而达到消除异常样本的目的。以图 4 为例，假定 $\varepsilon=0.5$，$k=3$，找到图 4 中箭头所指样本的 k 阶最近邻集合，如图所示，计算它的 Neighbors_Match$(x',k)=0<\varepsilon$，剔除箭头所指样本。

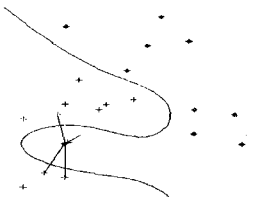

图 4　k 阶最近邻示意图

4　实验仿真

本实验分 P1 和 P2 两部分进行。第一部分（P1）是对双螺旋线的分类识别，分别使用传统支持向量机和本文所提出的改进方法进行测试，并给出它们对应的分类超平面效果图。图 5(a)代表原始双螺旋线分布图，图 5(b)与图 5(c)分别为约简前后的分类示意图。从图中可以发现，它们所对应的分类超平面几乎相同。图 5(d)是对约简后的样本添加噪声的分类示意图，噪声样本的存在使得分类超平面变得异常复杂。图 5(e)是剔除了噪声数据后的分类示意图。

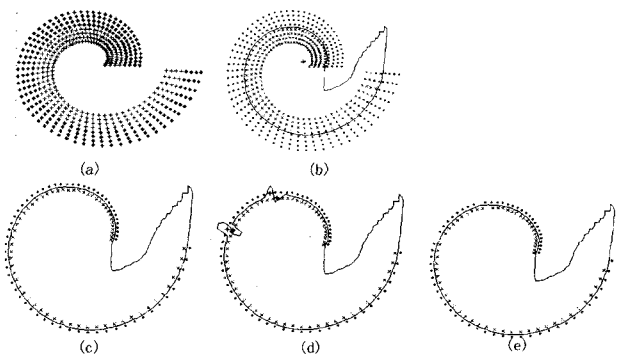

图 5　双螺旋线示意图

第二部分（P2）试验数据来自 UCI 机器学习知识库中的“Ionosphere，Wisconsin Diagnostic Breast Cancer，Cleveland Heart Disease”，数据集具体描述如表 2 所列。本部分试验分 3 组进行，每 1 组对原始样本集进行不同预处理。第 1 组对原始样本集不做任何处理，直接进行训练与测试，第 2 组首先对原始样本集进行属性约简，然后进行训练与测试，第 3 组是对样本集同时进行属性约简和样本选取预处理之后进行支持向量机的训练与测试。试验采用了 LIBSVM 软件包，核函数选用径向基 RBF 函数 $K(x,x_i)=\exp(-\|x-x_i\|^2/\sigma^2)$，参数 C 和 σ 使用网格搜索在[−10，10]获得。试验分别从特征个数、支持向量个数和分类精度 3 个方面对 3 组试验结果进行对比，具体内容如表 3 所列。表中支持向量和分类精度都是在 5 重交叉验证基础上取平均值。

表 2　数据集信息

Dataset Name	Samples	Features	Classes	Attribute Characteristics
Ionosphere	351	34	2	Integer，Real
Breast Cancer	569	32	2	Real
Heart Disease	303	75	2	Categorical，Integer，Real

（下转第 285 页）

从表1的数据可以看出，本文的算法在速度上比JM提供的标准算法有很大程度的提高，同时在信噪比和码率上最高只有0.03dB和0.12%的增加。另外，从表中的数据还可以知道，本文的算法在运动比较平缓的视频序列News上有更好的速度上的优势，这是因为在静止的视频序列上应用简单算法是有效的。

结束语 本文通过结合高精度运动估计算法和简单预测点运动估计算法的优势，提出了一种能在速度上与简单算法相当、在准确度上与高精度算法媲美的快速运动估计算法。本算法通过计算空间相邻运动向量差异(MVD)，自适应地判断采用何种运动估计方法。同时，提出了一种新的宏块运动估计顺序，其代替传统的宏块编码顺序作为新的运动估计的顺序，克服了位于图像边缘宏块运动估计的不准确性。根据实验表明，本文提出的快捷高精度运动估计方法，在相同码率和相同质量的情况下，能比原来的算法节省70%左右的搜索时间。

参 考 文 献

[1] Xu X, He Y. Improvements on Fast Motion Estimation Strategy for H. 264/AVC[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(3): 285-293

[2] Luo J, et al. Motion Estimation for Content Adaptive Video Compression[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2008, 18(7): 900-909

[3] Chen T-C, et al. Fast Algorithm and Architecture Design of Low-power Integer Motion Estimation for H. 264/AVC[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(5): 568-577

[4] Ahmad I, et al. A fast adaptive motion estimation algorithm[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(3): 420

[5] Lee Y G, Ra J B. Fast Motion Estimation Robust to Random Motions Based on a Distance Prediction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(7): 869-875

[6] Qaralleh E A A, Chang T S. Fast Variable Block Size Motion Estimation by Adaptive Early Termination[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2006, 16(8): 1021-1026

[7] Chen Z, et al. Fast integer-pel and fractional-pel motion estimation for H. 264/AVC[J]. Journal of Visual Communication and Image Representation, 2006, 17(2): 264

[8] Chao H, Lu J, Fisher P. Hybrid method for fast block based motion estimation in video coding[J]. Journal of Computational Information Systems, 2005, 1(2): 309

[9] Tourapis A M, Cheong H-Y, Topiwala P. Fast ME in the JM reference software, JVT-P026. doc. in Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q. 6) 16th Meeting[S]. Poznan, Poland, 2005

(上接第231页)

表3 不同方法的实验结果

	Group one			Group two			Group three		
	Features	SVs	Accuracy	Features	SVs	Accuracy	Features	SVs	Accuracy
Ionosphere	34	171	0.942	9	127	0.953	9	91	0.943
Breast Cancer	31	279	0.978	8	114	0.976	8	88	0.970
Heart Disease	14	122	0.921	5	83	0.927	5	69	0.922

结束语 本文针对支持向量机方法对于高维大规模数据无法直接处理和对异常样本敏感的问题，提出了一种基于邻域粗糙集模型的改进支持向量机，从属性约简和样本选取两个方面对样本数据进行预处理。最终试验结果证实本方法在提高支持向量机效率，改善支持向量机结构方面是可行的。由于在本方法中涉及了多个参数的选取，因此需要进一步研究各参数的选取和分析它们之间的关系。

参 考 文 献

[1] 李国正，王猛. 支持向量机导论[M]. 北京：电子工业出版社，2005

[2] Shin H, Cho S. Fast pattern selection for support vector classifiers[J]. Lecture Notes in Artificial Intelligence, 2003, 2637: 376-387

[3] Abe S, Inoue T, Fast training of support vector machine by extracting boundary data[C]// Proceeding of the International Conference on Artificial Neural Networks (ICANN), 2001: 308-313

[4] 安金龙，王正欧. 预抽取支持向量的支持向量机[J]. 计算机工程，2004，30(10)：9-11

[5] 李红莲，王春花，袁保宗. 一种改进的支持向量机NN-SVM[J]. 计算机学报，2003，26(8)：1015-1019

[6] Hu Q H, Yu D R, Xie Z X. Neighborhood Classifiers[J]. Expert System with Applications, 2008, 34(2): 866-876

[7] Hsu Chih-wei, Chang Chih-chung, Lin Chih-jen. A Practical Guide to Support Vector Classification[EB/OL]. http://home.eng. iastate. edu/~ julied/classes/ee547/Handouts/SVMuseguide. pdf

[8] Lin T Y. Granular Computing on binary relations I: data mining and neighborhood system[C]// Proc. of Rough Sets in Knowledge Discovery. Heidelberg, Germany: Physica-Verlag, 1998: 107-121

[9] Li Ye, Cai Yun-ze, Li Yuan-gui, et al. Rough Set Method for SVM Data Processing[C]// Proceeding of the 2004 IEEE Conference on Cybernetics and Intelligent System Singapore. 2004: 1039-1042

[10] 邓九英，杜启亮，毛宗源，等. 基于粗糙集与支持向量机的分类算法[J]. 华南理工大学学报，2008，36(5)：123-127

[11] 张建明，曾建武，谢磊，等. 基于粗糙集的支持向量机故障诊断[J]. 清华大学学报，2007，47(S2)：1774-1777

[12] 胡清华，赵辉，于达仁. 基于粗糙集的符号与数值属性的快速约简算法[C]// 第七届中国Rough集与软计算学术会议. 山西，太原，2007

[13] 王凯，张永祥，姚晓山，等. 支持向量机惩罚参数的自适应调整方法[J]. 计算机工程与应用，2008，44(26)：45-47

[14] Shin Hyunjung, Cho Sungzoon. Invariance of Neighborhood Relation Under Input Space to Feature Space Mapping[J]. Pattern Recognition Letters, 2004, 26(6): 707-718