◎数据库、信号与信息处理◎

处理非平衡数据的粒度SVM学习方法

徐 乾[1],王文剑[2],张文浩[1]
XU Qian[1],WANG Wenjian[2],ZHANG Wenhao[1]
1.山西大学 计算机与信息技术学院,太原 030006
2.山西大学 计算智能与中文信息处理教育部重点实验室,太原 030006
1.School of Computer and Information Technology,Shanxi University,Taiyuan 030006,China
2.Key Lab of Computational Intelligence & Chinese Information Processing of MoE,Shanxi University,Taiyuan 030006,China

XU Qian,WANG Wenjian,ZHANG Wenhao.Granular support vector machine approach used for imbalanced data.Computer Engineering and Applications,2011,47(24):97-99.

Abstract: Through the mining of multi-dimension association rules,Granular Computing(GrC) and Support Vector Machine (SVM) are efficiently amalgamated,and a Granular Support Vector Machine(GSVM) learning approach is proposed,namely AR-GSVM.For imbalanced datasets,AR-GSVM can not only reduce the complexity of the classifier,but also improve learning efficiency and generalization performance.Considering the data distribution consistence in the input space and kennel space,another granular SVM model on kennel space based on AR-GSVM is proposed,which is named as AR-KGSVM. AR-KGSVM can obtain better generalization performance comparing with AR-GSVM.The experimental results on UCI datasets demonstrate that the generalization performances of AR-GSVM and AR-KGSVM are superior to some most common used methods in dealing with imbalanced datasets.

Key words: support vector machine;granular computing;granular support vector machine;association rules;imbalanced data

摘 要: 通过多维关联规则挖掘,将粒度计算(Granular Computing,GrC)和支持向量机(Support Vector Machine,SVM)有效融合,提出一种粒度支持向量机(Granular SVM,GSVM)学习方法,称为AR-GSVM。该方法用于非平衡数据处理时,不仅可以有效降低分类器的复杂性,而且本质上可以进行并行计算以提高学习效率,同时提高分类器的泛化能力。考虑到保持数据在原始空间和特征空间的分布一致性,在AR-GSVM的基础上又提出核空间上的粒度支持向量机学习方法,称为AR-KGSVM,该方法具有更好的泛化性能。通过在UCI数据集上的实验表明:AR-GSVM和AR-KGSVM的泛化能力优于一些常用非平衡数据处理的方法。

关键词: 支持向量机;粒度计算;粒度支持向量机;关联规则;非平衡数据
DOI:10.3778/j.issn.1002-8331.2011.24.027 **文章编号:** 1002-8331(2011)24-0097-03 **文献标识码:** A **中图分类号:** TP301.6

1 引言

在许多数据分类应用中如医疗诊断、欺骗检测、垃圾邮件过滤等,经常会遇到数据中某一类样本在数量上远多于另一类的问题,这类问题通常称为非平衡数据处理问题。目前解决非平衡数据处理问题的主流方法有:上采样(增加少数类样本的数量)[1]、下采样(减少多数类样本的数量)[2]、代价敏感性学习(Cost Sensitive Learning,CSL)[3]、一类学习(One Class Learning,OCL)[4]、主动学习(Active Learning,AL)[5]等。这些方法虽有一定效果,但也存在一定的局限性,如上采样会增大数据量从而加重训练负担;下采样则极有可能消去对分类有用的样本;CSL方法需要调整敏感性参数cost值来达到分类性能最优,然而具体cost参数设置是个难题;OCL仅对少数类样本进行训练,适用于两类样本悬殊比较大的情况;AL方法选择离分类超平面较近的样本进行训练,从而降低了整个样本的正负类比例。因此,进一步开展非平衡数据处理的有效方法研究具有重要的应用价值。

粒度计算是当前智能信息处理领域中模拟人类思考问题和解决大规模复杂问题的新理论,而支持向量机在解决小样本、非线性及高维模式识别中表现出许多特有的优势,将二者有效地进行融合,可以将支持向量机这一优秀的机器学习方法更好地应用于实际问题的求解中。在粒度意义下探讨支持向量机学习机制,并明确粒度支持向量机含义的是由Y.C.Tang

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.60975035,No.71031006);教育部博士点基金(No.20091401110003);山西省自然科学基金重点项目(No.2009011017-2);山西省回国留学人员科研资助项目(No.2008-14)。
作者简介: 徐乾(1981—),男,博士生,主要研究方向:机器学习;王文剑(1968—),女,教授,博士生导师;张文浩(1983—),女,硕士生。
E-mail:xuqian@sxu.edu.cn
收稿日期: 2011-03-28;**修回日期:** 2011-05-16

于2004年提出的[6],其主要思想是首先构建粒度空间获得一系列信息粒,然后在每个信息粒上进行SVM学习,最后通过聚合信息粒上的信息获得最终的决策函数。目前关于粒度支持向量机模型的研究大体上有:SVM与粗糙集的结合[7-8];SVM与决策树的结合[9];SVM与聚类的结合[10-11];SVM与商空间的结合[12];SVM与关联规则的结合[13]等。

本文结合关联分类思想、粒度计算理论和传统SVM分类方法,提出了一种基于关联规则的粒度SVM(Association Rules based Granular SVM,AR-GSVM)学习算法,通过定义支持度和关联度两个度量获得规模与纯度合适的粒度层次,在此基础上进行训练样本的约简。在此基础上将样本由原始空间映射到高维空间,可得到基于关联规则的核粒度SVM(Association Rules based Kernel Granular SVM,AR-KGSVM)学习方法,粒划分和SVM训练都在高维空间中进行,从而保证了数据分布的一致性,进一步提升算法的泛化能力。通过实验表明:对非平衡数据的处理,AR-GSVM和AR-KGSVM算法明显优于其他几种传统方法。

2 基于多维关联规则挖掘的粒度支持向量机学习原理

2.1 基于关联规则的GSVM学习方法

基于关联规则的粒度支持向量机的基本思想是运用关联规则挖掘的方法,将样本空间划分为多个粒(子空间),即得到一系列纯粒(样本全部属于正类或负类)和混合粒(既包含正类样本也有负类样本),依次去掉关联度最高的纯粒,再在相应的混合粒中用SVM进行训练,循环执行直到分类准确率不再改进为止。

图1通过示例直观地表示出AR-GSVM与SVM分类效果的比较。从图中可以明显看到:SVM是在整个样本空间中分类,而AR-GSVM在$0\leq X\leq 3$和$6\leq X\leq 9$两个区域得到两个不需构建分类器的纯粒,只需对$3\leq X\leq 6$区间构成的混合粒进行分类,训练样本减少,超平面间隔增大。

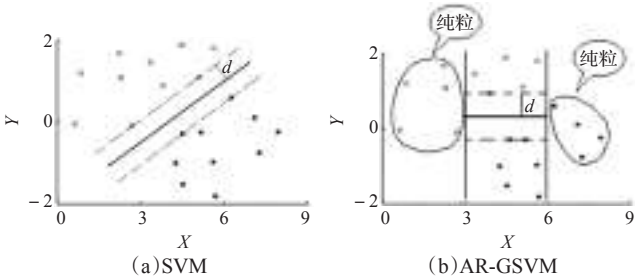

图1 SVM与AR-GSVM分类超平面的几何解释图

2.2 基于核空间的GSVM学习方法

目前,大部分基于粒度计算思想的GSVM模型都是在原空间进行粒划分,然后在核空间中进行训练。由于粒划分是在原空间中进行的,粒划分时非常可能丢失一些有用的信息,这样可能会导致数据在核空间分布与原空间分布不一致的问题,从而降低学习器的泛化能力。而基于核空间的GSVM则是将粒划分和学习统一在核空间中进行,与GSVM模型相比,KGSVM模型可克服数据分布不一致的问题,能获得更好的泛化性能。

3 AR-GSVM和AR-KGSVM学习算法基本思想

3.1 多维关联规则挖掘形成粒度

文献[13]虽给出了一种AR-GSVM模型,但粒划分只考虑了某一维属性与分类属性之间的关联性,忽略了属性间的关联,得到的关联规则有失合理性,本文则是通过定义关联度和支持度两个度量,挖掘多维关联规则来分割数据空间/特征空间以形成粒度,得到的关联规则更加合理。挖掘到的关联规则应满足左部是一个以上属性区间的合取,右部是一个分类属性,如:$A_{guan1}\wedge A_{guan2}\Rightarrow A_{cat}$,其中$A_{guan1}$和$A_{guan2}$是量化属性的区间,$A_{cat}$对于两类分类问题来说则是$y$=1或$y=-1$。

本文支持度support和关联度correlation分别定义如下:

$$support(A\Rightarrow B)=P(A\cup B)$$

$$correlation(A\Rightarrow B)=\frac{P(B|A)}{P(B|\bar{A})}$$

支持度反应的是符合规则的数据样本在整个数据集中占的百分比,关联度则反应了A对B的贡献率。关联度等于1,说明A的存在与否和B无关;关联度大于1,则说明A的出现对B有诱导作用。关联度越大,相关程度越高。

从粒度角度讲,当信息粒支持度和关联度大于阈值时,称作是纯粒,否则是混合粒。支持度的大小决定粒度的粗细;关联度越大则表明关联规则左部对右部的贡献率越大,得到的粒度纯度也就越高。因此,支持度和关联度两个参数的阈值设置需合理折中。

3.2 关联规则形成核粒度

在核空间中挖掘关联规则需要对距离的度量进行重新定义。

假设输入空间的样本$x_i\in\mathbf{R}^N, i=1,2,\cdots,l$,被非线性映射函数$\Phi$映射到某一特征空间$H$得到$\Phi(x_1),\Phi(x_2),\cdots,\Phi(x_l)$,那么输入空间的点积形式在特征空间用核函数表示为:

$$K(x_i,x_j)=(\Phi(x_i)\cdot\Phi(x_j)) \tag{1}$$

在核特征空间中欧式距离可以表示为:

$$d_{H(x,y)}=\sqrt{\|\Phi(x)-\Phi(y)\|^2=\Phi(x)\cdot\Phi(x)-2\Phi(x)\cdot\Phi(y)+\Phi(y)\cdot\Phi(y)} \tag{2}$$

一般情况下,非线性函数的表达式是未知的,所以结合式(1),式(2)可以写为:

$$d_{H(x,y)}=\sqrt{K(x,x)-2K(x,y)+K(y,y)} \tag{3}$$

上式即为欧式距离在核空间中的表达式。关联规则的挖掘也相应采用基于距离的挖掘方法。

3.3 AR-GSVM和AR-KGSVM算法

AR-GSVM算法的主要步骤如下:

输入 数据集l,关联规则支持度和关联度阈值,簇k的个数。

输出 分类正确率(或误分样本数)。

步骤1 将数据集l用聚类的方法划分成k个簇,然后把这k个簇投影到数值型属性所在的域以形成重叠的区间,最后这些重叠区间转化为布尔值。目标是将量化关联规则转化为布尔型关联规则的挖掘。

步骤2 使用关联规则挖掘算法(例如Apriori)挖掘形如"if $x_0<x<x_1$,then $y=1/y=-1$"的关联规则,并计算规则的支持度和关联度,将大于阈值的关联规则放置于集合AR中。

步骤3 取集合AR中关联度最大的关联规则,将规则对应下的样本(纯粒)赋相应类标签。

步骤4 使用交叉验证法找到最佳的核参数,对混合粒样本用SVM分类器进行训练。

步骤5 用构造的分类器模型预测未知类标签的测试样

本，得到相应的分类准确率 ar 。

步骤6 移除当前关联规则，返回执行步骤3，直到 AR 为空集。得到当前的分类准确率 $ar-cur$ ，如果 $ar-cur\geq ar$ ，则更新分类准确率 $ar=ar-cur$ ；否则，$ar=ar$ 。

步骤7 输出最好的分类结果。

AR-KGSVM是在上述算法基础上，选择核函数 k 将数据集 l 中的样本映射到特征空间中，在核空间中依距离度量式(3)将数据集 l 聚成 k 个类，其余步骤相同。

因此，AR-GSVM算法略去了不需参加训练的纯粒，只对混合粒样本做训练。而对于AR-KGSVM来说，除了上述AR-GSVM的优点外，AR-KGSVM是将样本映射到高维空间后，再作粒划分，粒划分和数据训练是在同一空间进行，保证了数据分布的一致性，泛化能力进一步提升。因而AR-KGSVM表现出更优越的特性。

4　实验结果与分析

4.1　分类器性能评价的标准

本文用以下参数来评价分类器的分类和预测性能。

(1)敏感性(sensitivity)：$sensitivity=\frac{TP}{TP+FN}$ ，指的是分类器正确预测的正类样本比例，实际上sensitivity代表了对正类的分类精度。

(2)特效性(specificity)：$specificity=\frac{TN}{TN+FP}$ ，指的是分类器正确预测的负类样本比例，实际上specificity代表了对负类的分类精度。

(3)几何均值(geometry means，G_means) $G\_means=\sqrt{sensitivity*specificity}$ ，这一指标是两类分类精度的几何均值，代表了整体的分类性能。

其中 TP、FP、TN、FN 的含义如表1所示。

表1　二分类问题的混合矩阵

实际样本数	预测为正类的样本数	预测为负类的样本数
实际的正类样本数	TP	FN
实际的负类样本数	FP	TN

4.2　实验结果与分析

将本文提出的算法AR-GSVM和AR-KGSVM与几种传统分类器(传统SVM、决策树-C4.5、神经网络-ART)和用于处理非平衡数据的两种主流方法(CSL SVM、AL SVM)进行了比较。

实验采用UCI数据集Wisconsin Breast、Banana、Thyroid，实验中训练集和测试集的设计情况如表2。图2~4为七种方法在不平衡程度逐渐变化的情况下各个指标的性能。

表2　实验中用到的非平衡数据集

数据集	实验分组	训练集			测试集	
		比例	正类	负类	正类	负类
Wisconsin Breast	1	1:3	100	300	50	150
	2	1:7	50	350		
	3	1:11	33	363		
Banana	1	1:3	150	450	100	300
	2	1:7	100	700		
	3	1:11	66	726		
Thyroid	1	1:3	400	1 200	200	600
	2	1:1 000	50	50 000		

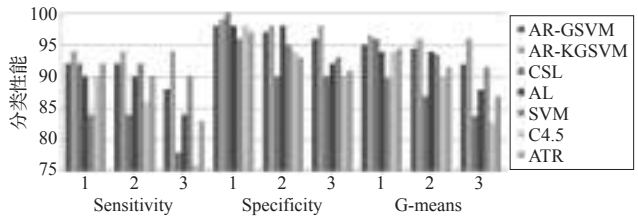

图2　不同比例下数据集Wisconsin Breast在三个指标上的分类性能比较

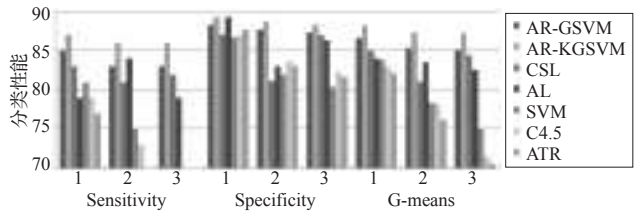

图3　不同比例下数据集Banana在三个指标上的分类性能比较

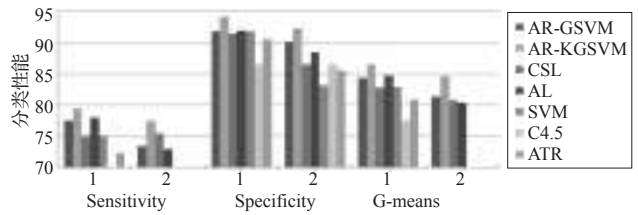

图4　不同比例下数据集Thyroid在三个指标上的分类性能比较

从图2~4可以看出：实验1中正负类样本略有不平衡，七种分类器在三个指标度量上都具有较好的分类性能，同等条件下AR-GSVM和AR-KGSVM略优于其余五种分类器。实验2、3中，正负类样本不平衡程度增大，SVM、C4.5和ART的指标sensitivity的值明显降低，导致G-means的值同样不高；CSL则随着样本不平衡程度增大，算法分类结果不稳定，收敛速度和算法精度不易协调；AL方法在上述两个样本集下分类结果较为稳定，但分类正确率不高，仅优于SVM、C4.5和ART三个传统分类方法。然而，AR-GSVM和AR-KGSVM在保证了较高的specificity的同时，sensitivity依然保持稳定。三个指标同时显示AR-GSVM和AR-KGSVM对正负类都具有良好的分类水平，其中，AR-KGSVM分类性能略优于AR-GSVM。

从图4还可以看出，样本正负类比例1:3的情况下，七种分类器均能对样本进行分类，但AR-GSVM和AR-KGSVM则表现出更优越的分类性能。然而在正负类样本严重失衡时(1:1 000)，CSL在不同参数下对正负类样本分类能力不稳定，AL、SVM、C4.5对正类样本的分类能力骤降，ART则基本失去对正类样本的分类能力，AR-GSVM和AR-KGSVM却保持对正负类样本都较好的分类性能。

5　结束语

提出一种基于多维关联规则挖掘的粒度支持向量机学习方法，在SVM与粒度的两种结合模式下分别设计算法AR-GSVM和AR-KGSVM。实验表明本文提出的方法比传统的SVM泛化能力强，且对非平衡数据的学习非常有效。本文仅限于对二分类问题的研究，如何将算法拓展到多类不平衡数据的分

(下转114页)