ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 3,May 2022,pp. 633－642
DOI:10. 16337/j. 1004-9037. 2022. 03. 013

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于粒计算的支持向量数据描述分类方法

方　宇[1]，曹雪梅[1]，杨　梅[1]，王　轩[2]，闵　帆[1]

(1. 西南石油大学计算机科学学院，成都 610500;2. 西南石油大学网络与信息化中心，成都 610500)

摘　要：分类学习效果与有限训练样本的分布情况密切相关。支持向量数据描述(Support vector data description，SVDD)作为单一边界求解模型，不能良好刻画数据实际分布特征，从而导致部分目标对象落在超球以外。为了提高其分类能力，本文提出一种基于粒计算的支持向量数据描述(Granular computing-driven SVDD，GrC-SVDD)分类方法，构造多粒度层次的属性集合以及相应的多粒度超球。首先通过邻域自信息对当前粒度层的属性集合重要度进行计算，然后选择最佳属性集合对上一粒度层未达到纯度阈值的超球再训练，直到所有超球满足条件或者属性耗尽。实验部分讨论了算法参数对分类性能的影响，并通过学习获得超参数。结果表明，与SVDD及流行的分类算法相比，本文方法具有较好的分类性能。

关键词：粒计算；支持向量数据描述；超球；邻域自信息；特征选择

中图分类号：TP181　　**文献标志码**：A

Granular Computing-Driven Support Vector Data Description Approach to Classification

FANG Yu[1]，CAO Xuemei[1]，YANG Mei[1]，WANG Xuan[2]，MIN Fan[1]

(1. School of Computer Science，Southwest Petroleum University，Chengdu 610500，China;2. Network and Information Center，Southwest Petroleum University，Chengdu 610500，China)

Abstract：The effect of classification learning is closely related to the distribution of limited training samples. Support vector data description (SVDD)，as a single boundary solution model，cannot well describe the actual distribution characteristics of the data，resulting in some target objects falling outside the hypersphere. To improve its classification ability，this paper proposes a granular computing-driven SVDD (GrC-SVDD) classification method to construct a multi-granularity levels attribute sets and the corresponding multi-granular hyperspheres. Firstly，the importance of the attribute within the current granularity level is calculated through the neighborhood self-information. Secondly，the best attribute set is then chosen to retrain the hyperspheres that did not achieve the purity criterion at the previous granularity level，and so on until all hyperspheres meet the conditions or the attributes are exhausted. The experimental section discusses the effect of parameters on classification performance and learns hyperparameters. The experimental results show that GrC-SVDD has better classification performance compared with SVDD and popular classification methods.

基金项目：国家自然科学基金(62006200)；四川省青年科技创新团队项目(2019JDTD0017)；西南石油大学研究生全英文课程建设项目(2020QY04)。
收稿日期：2021-04-03；**修订日期**：2021-11-08

Key words: granular computing; support vector data description; hyperspheres; neighborhood self-information; feature selection

引　　言

分类问题作为一项重要的任务,在机器学习、模式识别和数据挖掘领域有着广泛的应用[1-2]。在分类问题的模型构造上面,通常分为多分类方法和单分类方法,支持向量数据描述(Support vector data description,SVDD)[3]作为单分类方法的研究热点,其理论源于支持向量机[4](Support vector machine,SVM),特别适用于分类[5]、异常检测[6]、机械性能评估[7]、质量过程监测[8]和医学诊断[9]等领域。传统的SVDD是通过在高维特征空间中训练最小超球,以此确定目标对象决策边界,从而达到分类和异常检测的目的。优化SVDD的决策边界是SVDD理论的研究重点,众多学者相继提出了各种优化方法。Tax等[10]使用核主成分分析对样本做预处理,保证所有维的数据分布方差一致,从而得到更优的分类决策边界;Hoffmman等[11]使用特征空间映射与主成分投影的差作为新量度,使得SVDD的性能有显著的提高。但计算核PCA的代价、提取核矩阵特征值的代价均较高;Sadeghi等[12]使用模糊集验证度加权,使决策边界侧重关系密切的数据。Cha等[13]使用密度加权,以增加对密集数据误分类的惩罚;在算法层面,Peng等[14]提出了E-SVDD算法提高对象预测速度。Kim等[15]提出了深度学习神经网络模型用于改善SVDD的过拟合问题。在多分类方面,Huang等[16]提出了SVDD两类分类器的算法,Zhu等[17]提出了基于推广能力测度的多类SVDD模式识别方法。

以上针对SVDD的优化方法均是在一个粒度下求解的分类决策边界,这会存在目标对象未被划分的情况。粒计算[18]是一种高效、可扩展的计算方法,其粒化思想较好地解释了在分析和处理事务中采用自上而下、逐步求精的策略。这种策略在复杂分类问题中的应用优势尤为明显[19]。因此,本文基于粒计算思想,提出一种新分类方法。在该方法中,选择合适的属性重要度排序方法是构建粒的关键。Pawlak提出的粗糙集理论是一种处理不确定、不精确和不完备数据的有效手段[20]。众多研究者基于它提出了启发式特征选择算法。苗夺谦等[21]提出利用互信息衡量属性重要度的方法。王国胤等[22]设计了一种利用条件熵作为启发式信息函数的特征选择算法。宋桂娟等[23]利用决策属性相对于条件属性的条件熵和互信息的概念,提出了基于信息熵属性约简算法。然而,现有的大多数特征选择方法均是基于决策下近似构造的,这样会导致算法只考虑样本决策一致性的分类信息,而忽略决策分歧的样本提供的分类信息。因此,王长忠等[24]提出了综合考虑决策上下近似的邻域自信息,可以更准确地评估属性重要度。

1　支持向量数据描述

SVDD的目的是找到目标数据的最佳数据描述[3],即找到包含尽可能多的目标数据的最小超球。假设有数据集$\{x_i,\ i=1,\ \cdots,\ l\}$,其中$l$为目标类对象个数,SVDD的目标函数为

$$\min R^2+C\sum_i\xi_i$$

$$\text{s.t. } |x_i-a|^2\leqslant R^2+\xi_i,\quad \xi_i\geqslant 0\ \ \forall_i \tag{1}$$

式中:R为超球半径;C为平衡超球体积和球外目标数据数量之间的参数;ξ_i为松弛变量;a为超球中心。如图1所示,绿色点为目标类数据,黄色点为非目标数据,虚线圆则表示所求最小超球。

为了解决这些约束条件下的优化问题,构造了如下拉格朗日函数

$$L\left(R,a,\alpha_i,\gamma_i,\xi_i\right)=R^2+C\sum_i\xi_i-\sum_i\alpha_i\left\{R^2+\xi_i-\left(\|x_i\|^2-2a\cdot x_i+\|a\|^2\right)\right\}-\sum_i\gamma_i\xi_i \tag{2}$$

式中:$\alpha_i \geqslant 0, \gamma_i \geqslant 0$是拉格朗日乘数。为求拉格朗日函数的驻点,将偏导数设为0。

$$\frac{\partial L}{\partial R}=0:\ 2R-2R\sum_{i=1}^{l}\alpha_i=0\ \therefore \sum_{i=1}^{l}\alpha_i=1 \tag{3}$$

$$\frac{\partial L}{\partial a}=0:\ 2a-2\sum_{i=1}^{l}\alpha_i x_i=0\ \therefore a=\sum_{i=1}^{l}\alpha_i x_i \tag{4}$$

$$\frac{\partial L}{\partial \xi_i}=0\ \therefore C-\alpha_i-\gamma_i=0\ \forall_i \tag{5}$$

将新的约束式(3,5)代入拉格朗日函数式(2)中,得到如下新的目标函数

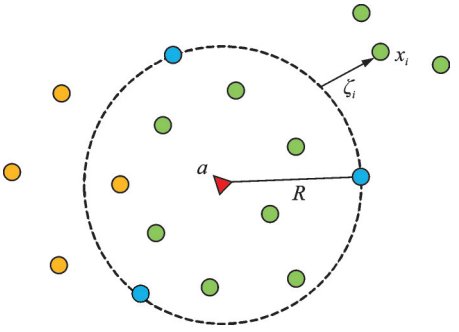

图1　特征空间中的SVDD

Fig. 1　SVDD in feature space

$$\begin{cases}\max \sum_{i=1}^{l}\alpha_i\langle x_i,x_i\rangle-\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j\langle x_i,x_j\rangle \\ \text{s.t. } 0\leqslant\alpha_i\leqslant C\quad i=1,2,\cdots,l \\ \sum_{i=1}^{l}\alpha_i=1\end{cases} \tag{6}$$

由于新的目标函数与向量之间的内积有关,所以可以用满足Mercer定理[3]的核函数$k(x_i,x_i)$代替内积$\langle x_i,x_j\rangle$,搜索最优数据描述就等价于

$$\begin{cases}\max \sum_{i=1}^{l}\alpha_i K(x_i,x_i)-\sum_{i=1}^{1}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i,x_j) \\ \text{s.t. } 0\leqslant\alpha_i\leqslant C\quad i=1,2,\cdots,l \\ \sum_{i=1}^{l}\alpha_i=1\end{cases} \tag{7}$$

因此,可以根据式(7)计算每个α_i的值;再依据任何非零α_i的线性组合,即支持向量(Support vectors,SVs)和式(4)来求出最优超球的中心,和式(8)求出超球半径,以此确定最佳的边界描述。SVs如图1边界蓝色点所示。

$$R^2=K(x_k\cdot x_k)-2\sum_i\alpha_i K(x_i\cdot x_k)+\sum_{i,j}\alpha_i\alpha_j K(x_i\cdot x_j) \tag{8}$$

为了判定测试样本z的类别,即z是否在超球体内,需计算样本z到超球中心a的距离,当其距离小于等于超球半径时,即

$$\|z-a\|^2=K(z\cdot z)-2\sum_i\alpha_i K(z\cdot x_k)+\sum_{i,j}\alpha_i\alpha_j K(x_i\cdot x_j)\leqslant R^2 \tag{9}$$

判定样本z为超球内样本;否则为超球外样本。

2　SVDD粒计算模型

SVDD粒计算模型分为两个步骤:首先利用邻域自信息计算粒层属性集合,其次根据粒层属性集合构造多粒度超球。本节详细介绍邻域自信息,并给出GrC-SVDD模型构造。

2.1　邻域自信息

邻域粗糙集是处理数据挖掘不确定性的有效方法之一[25],邻域自信息是利用邻域粗糙集理论中的上下近似概念构造的不确定性测度[24],能同时考虑决策上下近似包含的分类信息,准确地刻画属性重要度。

定义1[26]　决策表是一个五元组

$$S=(U,C,D,\{V_a|a\in At\},\{I_a|a\in At\}) \tag{10}$$

式中：U为一个非空的有限对象集合；C为条件属性集合；D为决策属性集合；At为属性的非空有限集合；V_a为每个属性$a\in At$的值的集合；I_a为属性$a\in At$的信息函数。

定义 2[24] 给定决策表S，$B\subseteq At$，$d_B(x,y)$为对象x,y在属性子集B上的欧氏距离，δ为邻域粒度，邻域关系R_B^δ定义为

$$R_B^\delta=\left\{(x,y)\in U\times U:|d_B(x,y)\leqslant\delta\right\} \tag{11}$$

定义 3[24] 给定决策表S，$B\subseteq At$，对于任意的$x\in U$，邻域类$[x]_B^\delta$的定义为

$$[x]_B^\delta=\left\{y\in U:(x,y)\in R_B^\delta\right\} \tag{12}$$

定义 4[24] 给定决策表S，$B\subseteq At$，$x\in U$，R_B^δ为由B诱导的U上邻域关系，对于任意$X\subseteq U$，X的下近似和上近似定义为

$$\underline{R}_B^\delta(X)=\left\{x\in U:[x]_B^\delta\subseteq X\right\} \tag{13}$$

$$\overline{R}_B^\delta(X)=\left\{x\in U:[x]_B^\delta\cap X\neq\varnothing\right\} \tag{14}$$

定义 5[24] 给定决策表S，$U/D=\{E_1,E_2,\cdots,E_r\}$是对象集$U$被决策$D$划分成$r$个分明的决策类，$B\subseteq At$，$R_B^\delta$是由$B$诱导的$U$上邻域关系。确信决策指标$\mathrm{cert}_B(E_k)$和可能决策指标$\mathrm{poss}_B(E_k)$定义为

$$\mathrm{cert}_B(E_k)=\left|\underline{R}_B^\delta(E_k)\right| \tag{15}$$

$$\mathrm{poss}_B(E_k)=\left|\overline{R}_B^\delta(E_k)\right| \tag{16}$$

确信决策指标用决策的下近似的基数刻画，表示分类一致的样本数量。可能决策指标用决策的上近似的基数刻画，表示可能属于该决策类的样本数量。

定义 6[24] 给定决策表S，$U/D=\{E_1,E_2,\cdots,E_r\}$，$B\subseteq At$，决策自信息$I_B(E_k)$和决策信息系统的决策自信息$I_B(D)$定义为

$$I_B(E_k)=\left(1-\frac{\mathrm{cert}_B(E_k)}{\mathrm{poss}_B(E_k)}\right)\lg\frac{\mathrm{cert}_B(E_k)}{\mathrm{poss}_B(E_k)} \tag{17}$$

$$I_B(D)=\sum_{k=1}^{r}I_B(E_k) \tag{18}$$

例 1 表1为iris数据集决策表S，其中$At=\{a_1,a_2,a_3,a_4\}$为属性集，$U=\{x_1,x_2,\cdots,x_n\}$为对象集，$D=\{1,2,3\}$为决策属性集，通过决策属性D将论域U化成3个等价类$E_1=\{x_1,x_2,\cdots,x_i\}$，$E_2=\{x_{i+1},x_{i+2},\cdots,x_l\}$和$E_2=\{x_{l+1},x_{l+2},\cdots,x_n\}$。

设定邻域粒度$\delta=0.33$，首先依据式(15,16)，计算等价类E_1,E_2,E_3在a_1,a_2,a_3,a_4上的确信决策指标和可能决策指标。其次，根据式(18)计算a_1,a_2,a_3,a_4的决策自信息，得$I_{a_1}(D)=34.54$，$I_{a_2}(D)=34.54$，$I_{a_3}(D)=24.48$，$I_{a_4}(D)=23.46$。自信息最小的属性更为重要，所以a_4相比其他3个属性更有利于目标对象边界描述。

2.2 GrC-SVDD模型

对信息粒化合理的解释是GrC-SVDD多粒度空间构建的前提。信息粒代表一个U的一个对象子集，它描述了一个系统或问题的子部分。通过聚集相同粒度的信息粒，可以得到一个系统或问题的整体描述[27]。在GrC-SVDD中，粒超球即为信息粒。GrC-SVDD粒化过程可解释为对原始数据在粒层属性集合上求解粒超球的过程。同时，小的粒超球是由大的粒超球细化而来。

定义 7 给定决策表S，$A\subseteq C$，粒超球GB定义为

$$GB=\{x_i|\left(x_i-a\right)^{|A|}\leqslant R^{|A|},x_i\in U\} \tag{19}$$

式中:a和R为粒超球中心和半径,分别由式(4,8)计算所得,$|A|$为属性维度。

定义8　给定决策表S,假设有$n+1,n\geqslant 1$层粒度,在决策表S上的粒化定义为

$$G=\{g\left(A_i\right)|A_i\subset C\} \tag{20}$$

式中:$g\left(A_i\right)$为某一特定相同粒度的粒超球集合;$A_i(1\leqslant i\leqslant n)$为条件属性的子集,且满足条件$A_1\subset A_2\subset\cdots\subset A_n\subset A_{n+1}$。

定义9　给定一个粒超球GB,其纯度p定义为

$$p=(t-n)/o \tag{21}$$

式中:o为该粒超球所有对象个数;n为该粒超球中非目标对象个数,即误包含对象个数。

算法设计如算法1所示。以表1为例,第1轮先根据邻域自信息计算每个属性的重要度,然后选出相对最重要的属性作为当前粒层的属性集合$F_1=\{a_4\}$,分别把等价类E_1,E_2,E_3视为目标数据,在F_1上对其进行SVDD训练,得到每个等价类的粒超球。计算每个粒超球的纯度p,将其与设定的纯度阈值对比,看其是否达到纯度要求,若该粒超球的纯度达到阈值,则其训练结束。如未达到阈值,则需要增添信息,使边界刻画更加清晰。此时在剩余的属性中选择出与a_4配合最好的1个属性。即$I_{\{a_4,a_i\}}(D)(i=1,2,3)$最小的集合。经过计算,$I_{\{a_4,a_3\}}(D)$最小,则当前粒层属性集合$F_2=\{a_4,a_3\}$。利用$F_2$对上一粒层未达到纯度阈值的粒超球进行同样的SVDD训练和判断,直到所有细化的粒超球都达到纯度阈值或者属性集合$F=C$时,训练结束。此时,会得到若干达到纯度阈值的粒超球。当有测试样本时,根据式(9)计算样本离每个粒超球的距离,预测其类别与距离最近粒超球一致。

表1　决策表

Table 1　Example of decision table

U	a_1	a_2	a_3	a_4	D
x_1	5.1	3.5	1.4	0.2	1
x_2	4.6	3.1	1.5	0.2	1
⋮	⋮	⋮	⋮	⋮	⋮
x_i	5.7	4.4	1.5	0.4	1
x_{i+1}	7.0	3.2	4.7	1.4	2
x_{i+2}	6.3	3.3	4.7	1.6	2
⋮	⋮	⋮	⋮	⋮	⋮
x_l	6.0	2.9	4.3	1.3	2
x_{l+1}	6.1	2.6	5.6	2.2	3
x_{l+2}	6.9	3.1	5.6	2.4	3
⋮	⋮	⋮	⋮	⋮	⋮
x_n	6.7	3.0	5.2	2.3	3

算法1　支持向量数据描述的粒计算分类算法

输入:决策表S,邻域粒度参数δ,平衡参数C,高斯核带宽σ,纯度阈值ε。

输出:粒超球集合leaf。

(1) 初始化:red$=\varnothing$,$B=C-$red, $L=U$,leaf$=\varnothing$;/*red存放当前特征子集,B存放剩余特征,leaf存放满足条件的粒超球,L存放每个粒层粒超球集合*/

(2) while red$\neq C$

(3) $T=\varnothing$;

(4) for each $a_i\in B$

(5) $T\leftarrow$ red$\cup\{a_i\}$;

(6) 根据式(18)计算T的决策自信息$I_T(D)$;

(7) end for

(8) 找到使$I_T(D)$值最小的属性a_l;

(9) red$\leftarrow$ red$\cup\{a_l\}$,$B\leftarrow B-$ red;

(10) 对L求粒超球集合$g(\mathrm{red})$,根据式(22)计算每个粒超球GB的纯度p;

(11) $L=g(\mathrm{red})$

(12) 遍历$g(\mathrm{red})$中每个粒超球GB_i

(13) if $p_i>\varepsilon$

(14) $L\leftarrow L-GB_i$,跳转至步骤3;

(15) else

(16) $\mathrm{leaf}=\mathrm{leaf}\cup GB_i$

(17) end if

(18) end for

(19) end while

(20) return leaf

图2直观地描述了GrC-SVDD整个训练过程,其中黄、红、绿三种颜色对应的点分别表示类别1、类别2和类别3的样本,黑色虚线为粒超球决策边界,a_i和R_i分别是每个粒超球的中心和半径,红色虚线表示粒层间的映射。对于任意数据集U,认为其最开始为粗粒度状态。利用邻域自信息选出粒层1的属性集合F_1作为其特征空间,在这个特征空间上,对该数据集的r个类别的样本分别进行SVDD训练,最后得到r个粒超球。图2中的U即为3个类别,最后得到GB_1,GB_2和GB_3三个粒超球,计算每个粒超球的纯度,发现粒超球GB_1达到了纯度阈值,不再继续训练。这表明在特征空间F_1中,类别1在决策边界完全可分。粒超球GB_2和GB_3未达到纯度阈值,这表明在特征空间F_1中,类别2和类别3的决策边界描述还需要增添信息。因此,将其映射到特征空间F_2中,再次训练。此时,GB_2和GB_3中只包含类别2和类别3,只需分别对GB_2和GB_3中的类别2和类别3进行SVDD训练。得到粒超球GB_4,GB_5,GB_6和GB_7,计算其纯度,发现在特征空间F_2中,类别2和类别3在决策边界上已经完全可分,即达到纯度阈值,整个训练结束。当有测试样本时,计算其到粒超球GB_1,GB_4,GB_5,GB_6和GB_7的距离,预测其类别为最近粒超球的类别。

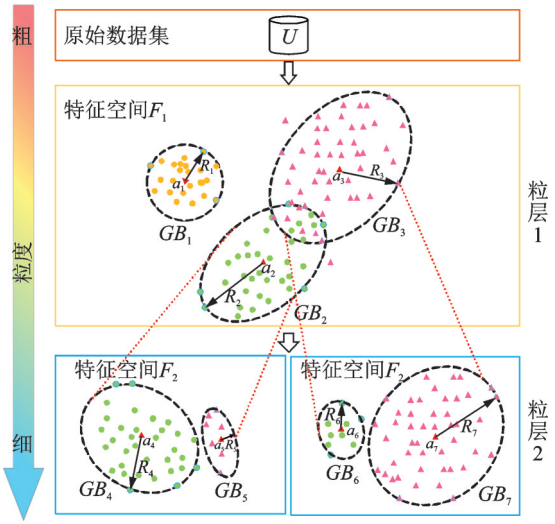

图2 GrC-SVDD粒化示意图

Fig. 2 GrC-SVDD granulation diagram

3 实验分析

实验选用10个UCI数据集运用于算法的验证,数据具体信息见表2前4列。实验采用十折交叉验证,依据精度的大小来评估GrC-SVDD与对比算法的分类性能,精度定义为

$$\mathrm{accuracy}=(N-e)/N\times 100\% \tag{19}$$

式中:N为总数据对象个数;e为误分类个数。为了获得最佳的实验结果,邻域半径δ和纯度阈值ε的设置至关重要。因此,通过讨论参数$\delta\in[0,1]$,步长为0.01和$\varepsilon\in[0.8,1]$,步长为0.1,找到每个数据集达到最佳精度时的参数取值。以iris数据集为例,图3、4分别显示了GrC-SVDD随邻域半径变化以及纯度阈值变化的分类精度。以此方法找到所有数据集的最优参数设置,详细取值见表2后两列。平衡参

数C设置为1,高斯核带宽$\sigma=1/(n\times t)$,其中n为目标对象个数,t为目标对象标准差。实验环境:操作系统,Windows10,CPU为AMD4800H,编译环境为Python3.8。

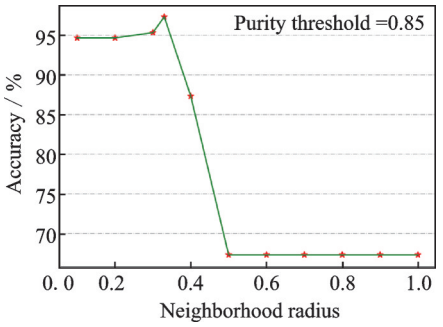

图3　iris分类精度随邻域半径的变化曲线

Fig. 3　Variation of classification accuracies with neighborhood radius

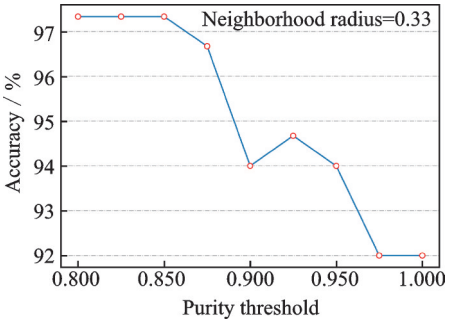

图4　iris分类精度随纯度阈值的变化曲线

Fig. 4　Variation of classification accuracies with purity threshold

表2　数据集信息

Table 2　Information of datasets

U	N	C	D	ε	δ
iris	150	4	3	0.85	0.33
cancer	689	9	2	0.96	0.12
seeds	210	7	3	0.98	0.23
appendicitis	106	7	2	0.94	0.65
breast	699	9	2	0.90	0.60
australian	690	14	2	0.85	0.50
vote	435	16	2	0.96	0.40
glass	214	10	7	0.95	0.65
crx	690	15	2	0.90	0.35
wine	170	13	3	0.96	0.30

GrC-SVDD在粒化的过程中,仅选择了部分属性用于训练,其训练平均使用属性个数如图5所示。Original表示数据集原始属性个数,同时是对比算法训练使用的属性个数。橙色表示GrC-SVDD训练所使用的平均属性个数。图中橙色柱条明显矮于蓝色柱条,即GrC-SVDD训练使用属性个数明显少于对比算法。特别是在glass,iris以及wine数据集上,GrC-SVDD仅用了少部分属性进行训练。

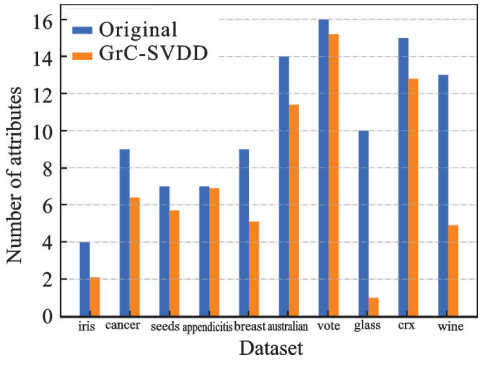

图5　原始属性个数和GrC-SVDD使用属性个数对比

Fig. 5　Comparison of the number of original attributes and the number of attributes used by GrC-SVDD

实验对比了5种分类算法:k近邻(k-nearest neighbor,kNN)[28]、决策树(Decision tree C4.8,J48)[29]、贝叶斯(Naive Bayesian,NB)[30]、逻辑回归(Logistic regression,LGR)[31]以及SVDD。实验结果如表3所示,加粗的数字表示与其他算法相比,在对应数据集上分类效果相对最优。加减的角标为十折交叉验证的精度标准差,表示在不同训练集上精度的波动,即模型的泛化能力。相对比之下,GrC-SVDD分类精度6次

达到最佳,其中在iris数据集上精度达到97.33%,明显优于其他分类算法。在这10个数据集上,GrC-SVDD精度Meanrank为2,位居第一。平均标准差相对最小,即精度波动最小。实验结果充分表明,GrC-SVDD不仅只使用了少量属性训练,而且分类性能较好,模型泛化能力也较强,适应大多数数据集。

表3 算法精度对比

Table 3 Accuracy comparison with other algorithms

Dataset	*k*NN	NB	J48	LGR	SVDD	GrC-SVDD
iris	$95.33_{\pm 6.70}$	$95.33_{\pm 5.21}$	$94.00_{\pm 6.29}$	$92.67_{\pm 4.67}$	$72.00_{\pm 10.24}$	$97.33_{\pm 4.42}$
cancer	$96.76_{\pm 1.95}$	$96.32_{\pm 2.65}$	$94.26_{\pm 2.97}$	$96.62_{\pm 2.38}$	$92.79_{\pm 3.92}$	$96.88_{\pm 2.85}$
seeds	$93.33_{\pm 5.71}$	$89.52_{\pm 6.67}$	$92.86_{\pm 5.32}$	$92.86_{\pm 7.75}$	$78.57_{\pm 10.49}$	$91.90_{\pm 4.15}$
appendicitis	$85.00_{\pm 8.06}$	$85.00_{\pm 8.06}$	$80.00_{\pm 10.00}$	$86.00_{\pm 11.14}$	$64.00_{\pm 11.14}$	$87.00_{\pm 6.40}$
breast	$96.38_{\pm 2.53}$	$95.94_{\pm 2.13}$	$94.64_{\pm 3.43}$	$95.94_{\pm 2.81}$	$91.74_{\pm 3.73}$	$96.38_{\pm 2.17}$
australian	$83.77_{\pm 5.01}$	$80.00_{\pm 3.36}$	$81.30_{\pm 3.00}$	$86.38_{\pm 5.03}$	$73.19_{\pm 5.66}$	$87.91_{\pm 4.27}$
vote	$91.86_{\pm 4.79}$	$93.72_{\pm 5.10}$	$94.42_{\pm 3.63}$	$93.95_{\pm 4.67}$	$76.28_{\pm 4.00}$	$92.79_{\pm 4.70}$
glass	$89.52_{\pm 7.32}$	$83.33_{\pm 8.84}$	$98.10_{\pm 2.33}$	$54.29_{\pm 8.02}$	$69.05_{\pm 12.46}$	$95.24_{\pm 4.26}$
crx	$84.93_{\pm 5.77}$	$79.71_{\pm 4.25}$	$80.43_{\pm 3.85}$	$85.94_{\pm 3.55}$	$61.30_{\pm 4.94}$	$85.98_{\pm 2.43}$
wine	$95.88_{\pm 4.59}$	$97.06_{\pm 2.94}$	$88.24_{\pm 9.11}$	$98.82_{\pm 2.35}$	$82.94_{\pm 10.00}$	$94.71_{\pm 4.89}$
Meanrank	2.8	3.6	3.5	3.0	5.8	2.0

4 结束语

针对SVDD不能很好刻画数据边界的问题,提出了基于粒计算的支持向量数据描述分类方法,即GrC-SVDD,通过实验分析得到如下结论:(1) 为了提高SVDD在单一决策边界上的分类能力,本文通过构造了多粒度层次属性集合,将单粒度超球扩展为了多粒度超球,使原本在单一决策边界上不可区分的样本在多粒度超球决策边界上变得可分。(2) 在粒化过程中,选择了有利于边界描述的属性集合,使得本文方法能在少量属性上做出快速决策。不仅分类效果较好,而且模型泛化能力较强。(3) 实验部分讨论了邻域半径δ和纯度阈值ε对分类结果的影响,并给出了取值建议。下一步研究工作:(1) 本文是以一定步长改变参数值,通过观察实验结果来确定最佳参数。若步长设置不合理以及取值区间过大或过小,则将需要较多时间来得到最佳结果。接下来的工作将研究更加高效的参数寻优方法,以进一步提高模型的分类性能。(2) 本文方法默认模型输入是数值型数据,但在大数据中,数据具有维数高和特征形式多样化的特点。如何利用GrC-SVDD对混合型数据进行复杂分类,这也将是下一步的研究重点。

参考文献:

[1] GU B, SUN X M, SHENG V S. Structural minimax probability machine[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(7): 1646-1656.

[2] ZHANG S C, LI X L, ZONG M, et al. Efficient kNN classification with different numbers of nearest neighbors[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1774-1785.

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

[4] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[5] 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.

[6] LIU Y H, LIU Y C, CHEN Y J. Fast support vector data descriptions for novelty detection[J]. IEEE Transactions on Neural

Networks, 2010, 21(8): 1296-1313.

[7] PAN Y, CHEN J, GUO L. Robust bearing performance degradation assessment method based on improved wavelet packet-support vector data description[J]. Mechanical Systems and Signal Processing, 2009, 23(3): 669-681.

[8] GE Z, SONG Z. Bagging support vector data description model for batch process monitoring[J]. Journal of Process Control, 2013, 23(8): 1090-1096.

[9] CAO J, ZHANG L, WANG B, et al. A fast gene selection method for multi-cancer classification using multiple support vector data description[J]. Journal of Biomedical Informatics, 2015, 53: 381-389.

[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.

[11] HOFFMANN H. Kernel PCA for novelty detection[J]. Pattern Recognition, 2007, 40(3): 863-874.

[12] SADEGHI R, HAMIDZADEH J. Automatic support vector data description[J]. Soft Computing, 2018, 22(1): 147-158.

[13] CHA M, KIM J S, BAEK J G. Density weighted support vector data description[J]. Expert Systems with Applications, 2014, 41(7): 3343-3350.

[14] PENG X, XU D. Efficient support vector data descriptions for novelty detection[J]. Neural Computing and Applications, 2012, 21(8): 2023-2032.

[15] KIM S, CHOI Y, LEE M. Deep learning with support vector data description[J]. Neurocomputing, 2015, 165: 111-117.

[16] HUANG G, CHEN H, ZHOU Z, et al. Two-class support vector data description[J]. Pattern Recognition, 2011, 44(2): 320-329.

[17] 朱孝开, 杨德贵. 基于推广能力测度的多类 SVDD 模式识别方法[J]. 电子学报, 2009, 37(3): 464-469.
ZHU Xiaokai, YANG Degui. Multi-class support vector domain description for pattern recognition based on a measure of expansibility[J]. Acta Electronica Sinica, 2009, 37(3): 464-469.

[18] ZADEH L A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets & Systems, 1997, 90: 111-127.

[19] XIA S, LIU Y, DING X, et al. Granular ball computing classifiers for efficient, scalable and robust learning[J]. Information Sciences, 2019, 483: 136-152.

[20] PAWLAK Z. Rough sets[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[21] 苗夺谦, 胡桂容. 知识约简的一种启发式算法[J]. 计算机研究与发展, 1999, 36(6): 42-45.
MIAO Duoqian, HU Guirong. A heuristic algorithm for reduction of knowledge[J]. Journal of Computer Research and Development, 1999, 36(6): 42-45.

[22] 王国胤, 于洪, 杨大春. 基于条件信息熵的决策表约简[J]. 计算机学报, 2002, 25(7): 759-766.
WANG Guoyin, YU Hong, YANG Dachun. Decision table reduction based on conditional information entropy[J]. Chinese Journal of Computers, 2002, 25(7): 759-766.

[23] 宋桂娟, 曲朝阳, 李翔坤, 等. 基于信息熵的粗糙集属性约简算法研究[J]. 微计算机信息, 2010 (18): 212-213.
SONG Guijuan, QU Zhaoyang, LI Xiangkun, et al. Research on attribute reduction algorithm in rough set based on information entropy[J]. Control & Automation, 2010 (18): 212-213.

[24] WANG C, HUANG Y, SHAO M, et al. Feature selection based on neighborhood self-information[J]. IEEE Transactions on Cybernetics, 2019, 50(9): 4031-4042.

[25] LIN T Y. Neighborhood systems and approximation in relational databases and knowledge bases [C]// Proceedings of the 4th International Symposium on Methodologies of Intelligent Systems. Northridge CA: California State University, 1988: 75-86.

[26] FANG Y, GAO C, YAO Y. Granularity-driven sequential three-way decisions: A cost-sensitive approach to classification[J]. Information Sciences, 2020, 507: 644-664.

[27] 方宇, 闵帆, 刘忠慧, 等. 序贯三支决策的代价敏感分类方法[J]. 南京大学学报:自然科学版, 2018, 54(1): 148-156.
FANG Yu, MIN Fan, LIU Zhonghui, et al. Sequential three-way decisions based cost-sensitive approach to classification[J]. Journal of Nanjing University: Natural Science, 2018, 54(1): 148-156.

[28] ALTMAN N S. An introduction to kernel and nearest-neighbor nonparametric regression[J]. The American Statistician, 1992,

46(3): 175-185.

[29] XIANG Z, ZHANG L. Research on an optimized C4.5 algorithm based on rough set theory[C]// Proceedings of 2012 International Conference on Management of e-Commerce and e-Government.[S.l.]: IEEE, 2012: 272-274.

[30] RISH I. An empirical study of the naive Bayes classifier[C]//Proceedings of IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.[S.l.]: IJCAI, 2001: 41-46.

[31] CAI Y D, FENG K Y, LU W C, et al. Using logitBoost classifier to predict protein structural classes[J]. Journal of Theoretical Biology, 2006, 238(1): 172-176.

作者简介:

方宇(1983-),男,副教授,研究方向:粗糙集、粒计算、三支决策等,E-mail: Fangyu@supu.edu.cn。

曹雪梅(1996-),女,硕士研究生,研究方向:三支决策、粒计算、机器学习等。

杨梅(1982-),女,副教授,研究方向:机器学习、多示例学习、推荐系统等。

王轩(1991-),男,助理实验师,研究方向:机器学习、计算机网络等。

闵帆(1973-),**通信作者**,男,教授,研究方向:机器学习、主动学习、粒计算等。

(编辑:刘彦东)