ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 1,Jan. 2017,pp. 46-53
DOI:10.16337/j.1004-9037.2017.01.005

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

一种高斯区间核 SVM 分类模型

王文剑[1,2]　祁晓博[1]　郭虎升[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)

摘　要: 区间型数据(Interval data, ID)是属性特征取值为区间的一类数据,针对区间型数据的分类问题,本文提出一种高斯区间核支持向量机分类模型(Support vector machine based on Gauss interval kernel, GIK_SVM)。该方法引入半宽因子,在区间型数据的中值与半宽度之间进行折中,并据此构造高斯区间核用以衡量两个区间型数据间的相似性,然后用 SVM 模型进行分类。在人造数据集和真实数据集上的实验结果表明,本文提出的算法对区间数据有更好的分类性能。

关键词: 区间型数据; 半宽因子; 区间核; GIK_SVM 模型
中图分类号: TP18　　**文献标识码:** A

Support Vector Machine Classification Model Based on Gauss Interval Kernel

Wang Wenjian[1,2], Qi Xiaobo[1], Guo Husheng[1]

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

Abstract: Interval data (ID) is a kind of data which the attribute values are the interval. Aiming at the classification problem of interval data, a support vector machine classification model based on Gauss interval kernel (GIK_SVM) is proposed. In the method, the half-width factor is introduced which makes a compromise between the median and the half width of interval data. Then, the Gauss interval kernel is constructed to measure the similarity between two interval data. SVM model is applied to classify the samples. Experiment results on artificial and real datasets demonstrate that the proposed GIK_SVM has a better classification performance for interval data.

Key words: interval data(ID); half-width factor; interval kernal; GIK_SVM model

引　言

随着互联网与信息技术的迅猛发展,数据的获取与使用逐渐便捷,不仅数据量每年都在飞速增长,数据的复杂性也日趋明显[1,2]。其中有一类数据与人们的生产、生活息息相关,如某一地区一段时间的气温变化、某一段时间内的交通流量和工业总产值增长率等,这类数据的特点是:每个属性特征的取值

基金项目: 国家自然科学基金(61673249, 61503229, 61273291)资助项目;山西省回国留学人员科研项目(2016-004)资助项目;山西省自然科学青年基金(2015021096)资助项目;山西省高等学校科技创新(2015110)资助项目。
收稿日期: 2016-11-30;**修订日期:** 2017-01-07

不确定,而是一个区间,这类数据称为区间型数据。这类数据的出现可能是由于属性值的多次测量、置信区间估计或取值范围有界等。相较于离散数据,区间型数据可以从全局把握数据对象的内在结构特征,更有利于揭示隐含在数据内部的规律。因此,区间型数据可以表示数据的不确定性和可变性,在决策支持中具有重要的应用价值。与离散数据(确定性数据)不同,目前关于区间型数据的处理方法主要有 3 大类。(1)模糊集方法[3],这类方法通过计算元素关于集合的隶属程度来近似描述不确定性,将每个区间离散化为一个确定值(通常是符号属性数据),再用传统方法进行处理。这类方法中隶属函数大多为专家凭经验给出,带有强烈的主观意志。(2)中值法[4],即用区间中值作为区间型数据的特殊点,再用传统方法进行处理。该方法只考虑了区间型数据的内部情况,丢失了区间大小这一相关信息。(3)采用上下边界值替代区间型数据[5-7],即将区间型数据离散化为两个确定性数值,再用传统方法进行处理。这类方法只用上下边界值进行计算,未考虑区间型数据的内部分布情况。因此又有学者提出改进方法,在上下界基础上考虑中值信息[8-10],这样不仅考虑到区间边界,还将内部分布一并融合进去,使区间型数据表示更加全面。文献[8]利用区间中值与宽度表示区间型数据,运用传统的回归方法分别对区间中值与区间半宽度生成回归方程,然后通过这两个方程对区间上下限进行预测。文献[9]用区间中值与宽度表示区间变量,在这两个独立的确定性变量上用对称线性回归模型进行预测。文献[10]提取区间值数据的区间中值与宽度,分别作为 Gauss 分布函数的期望和方差,用 Gauss 分布函数表示区间值数据并对其进行相似度量。

目前关于区间型数据的处理主要集中在聚类和回归分析中[4-11],分类问题的研究相对较少[12]。考虑到区间型数据的特点及支持向量机(Support vector machine, SVM)良好的泛化能力[13-14],本文提出一种高斯区间核 SVM 分类模型。该模型采用区间中值与半宽度表示区间型数据,设计了一个可调的半宽因子,并构造了高斯区间核,进而利用高斯区间核 SVM 模型对区间型数据进行分类。

1 基于高斯型区间核的 SVM 分类模型

1.1 高斯区间核

定义 1 (区间型样本):x_i 为一个区间型样本,记为 $x_i=(b_{i1},b_{i2},\cdots,b_{ik})$,其中,$b_{it}$($t\in[1,k]$)为一个区间型特征值,记为 $b_{it}=[b_{it}^l,b_{it}^u]$,$b_{it}^l,b_{it}^u\in\mathbf{R}$ 且 $b_{it}^l\leqslant b_{it}^u$,$b_{it}^l$ 称区间左边界,b_{it}^u 称区间右边界。

定义 2 (区间中值):称 $\bar{x}_i$ 为区间型样本 x_i 的中值,记为 $\bar{x}_i=(\bar{b}_{i1},\bar{b}_{i2},\cdots,\bar{b}_{ik})$,其中

$$\bar{b}_{it}=\frac{b_{it}^l+b_{it}^u}{2}\quad t\in[1,k] \tag{1}$$

定义 3 (区间半宽度):称 $\hat{x}_i$ 为区间型样本 x_i 的半宽度,记为 $\hat{x}_i=(\hat{b}_{i1},\hat{b}_{i2},\cdots,\hat{b}_{ik})$,其中

$$\hat{b}_{it}=\frac{b_{it}^u-b_{it}^l}{2}\quad t\in[1,k] \tag{2}$$

核方法是机器学习中解决非线性问题的一种重要技术,它通过核函数刻画特征空间向量间的内积,避开了非线性映射的显式表达,既保证了学习器的泛化性能又避免了维度灾难问题。本质上核还可以通过计算两个样本映射到高维空间之后的内积度量数据的相似性,但常用的线性核、高斯核和多项式核等一般不能直接衡量区间型数据的相似性。为此,本文构造了一种高斯区间核,用以衡量区间型数据的相似性。对于两个区间型样本 $x_i(b_{i1},b_{i2},\cdots,b_{ik})$ 和 $x_j(b_{j1},b_{j2},\cdots,b_{jk})$,高斯区间核构造为

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{3}$$

其中

$$\|x_i-x_j\|^2=(b_{i1}-b_{j1})^2+\cdots+(b_{ik}-b_{jk})^2 \tag{4}$$

$$(\boldsymbol{b}_{it}-\boldsymbol{b}_{jt})^2=\alpha_t(\bar{b}_{it}-\bar{b}_{jt})^2+(1-\alpha_t)(\hat{b}_{it}-\hat{b}_{jt})^2\quad t\in[1,k] \tag{5}$$

联立式(3～5),可得

$$K(x_i,x_j)=\exp\left(-\frac{\alpha_1(\bar{b}_{i1}-\bar{b}_{j1})^2+(1-\alpha_1)(\hat{b}_{i1}-\hat{b}_{j1})^2+\cdots+\alpha_k(\bar{b}_{ik}-\bar{b}_{jk})^2+(1-\alpha_k)(\hat{b}_{ik}-\hat{b}_{jk})^2}{2\sigma^2}\right)=$$

$$\exp\left(-\frac{\sum_{t=1}^{k}\left[\alpha_t(\bar{b}_{it}-\bar{b}_{jt})^2+(1-\alpha_t)(\hat{b}_{it}-\hat{b}_{jt})^2\right]}{2\sigma^2}\right) \tag{6}$$

式中:$\alpha_1,\cdots,\alpha_k\in[0,1]$,为区间型数据的半宽因子,半宽因子 α_t 使区间中值与区间半宽度对样本相似性度量达到有效折中。当 $\alpha=1$ 时,该方法只考虑区间中值;当 $\alpha=0$ 时,该方法只考虑区间半宽度;当 α 介于 0 和 1 之间时,用于折中区间中值与半宽度对区间型数据的相似性度量。很显然,式(6)满足核函数的条件,本文称之为高斯区间核。

1.2 算法的主要步骤

设区间型数据集 $T=\{(x_i,y_i)\}_{i=1}^{l}$,其中 x_i 为区间型样本,$y_i\in\{-1,1\}$ 为分类标识。本文提出 GIK_SVM 模型的主要思想是:首先求解区间型样本的区间中值与区间半宽度,根据式(6)构建高斯区间核矩阵,然后用高斯区间核 SVM 分类模型进行分类。GIK_SVM 算法的主要步骤如下。

算法 1 GIK_SVM 算法

输入:训练集 $Tr=\{(x_i,y_i)\}_{i=1}^{l}$,测试集 $Te=\{(x_j,y_j)\}_{j=1}^{l}$,高斯区间核的参数 σ,半宽因子 $\alpha_1,\cdots,\alpha_k$。

(1) 根据式(1,2)分别计算出 Tr 和 Te 上区间型样本的区间中值与区间半宽度;

(2) 根据式(6)在训练集 Tr 上计算高斯区间核矩阵;

(3) 根据所得高斯区间核矩阵建立 SVM 分类模型,并在 Te 上进行测试;

(4) 算法结束。

将本文算法与基于区间中值的 SVM 分类算法(Support vector machine based on interval median, IM_SVM)和基于区间边界值的 SVM 分类算法(Support vector machine based on interval boundary value, IBV_SVM)进行比较,其中,IM_SVM 算法只考虑区间型数据中值这一主要因素,IBV_SVM 算法只考虑区间的上下两个边界值,两种算法的主要步骤分别如下。

算法 2 IM_SVM 算法

输入:训练集 $Tr=\{(x_i,y_i)\}_{i=1}^{l}$,测试集 $Te=\{(x_j,y_j)\}_{j=1}^{l}$,高斯核的参数 σ。

(1) 根据式(1)分别计算出 Tr 和 Te 上区间型样本的区间中值,生成新的训练集 $Tr'=\{(\bar{x}_i,y_i)\}_{i=1}^{l}$ 和测试集 $Te'=\{(\bar{x}_i,y_i)\}_{i=1}^{l}$;

(2) 用传统高斯核在新的训练集 Tr' 上计算高斯核矩阵;

(3) 根据所得高斯核矩阵建立 SVM 分类模型,并在 Te' 上进行测试;

(4) 算法结束。

算法 3 IBV_SVM 算法

输入:训练集 $Tr=\{(x_i,y_i)\}_{i=1}^{l}$,测试集 $Te=\{(x_j,y_j)\}_{j=1}^{l}$,高斯核的参数 σ。

(1) 分别分离出 Tr 和 Te 上区间型样本的左右边界值,生成新的训练集 $Tr'=\{(x_i',y_i)\}_{i=1}^{l}$ 和测试集 $Te'=\{(x_i',y_i)\}_{i=1}^{l}$,其中,$Tr'$ 中每个 x_i' 用其左右边界值来代替,即 $x_i'=b_{it}^{l},b_{it}^{u}(t\in[1,k])$,这样,新的 x_i' 将比原始 x_i 多了 k 列;

(2) 用传统高斯核在新的训练集 Tr' 上计算高斯核矩阵;

(3) 根据所得高斯核矩阵建立 SVM 分类模型,并在 Te' 上进行测试;

(4) 算法结束。

标准 SVM 模型的时间复杂度为 $O(n^2)$,其中,n 为参与训练样本集的规模。假设数据维数为 d,IM_SVM 算法只取中值做特殊点,维数仍为 d,其复杂度为 dn^2;IBV_SVM 算法取两个边界值做特殊点,维数为 $2d$,时间复杂度为 $2dn^2$;GIK_SVM 取区间中值与半宽度两个值做特殊点,由于高斯区间核将区间中值与半宽度联系为一个整体,此时维数仍为 d,时间复杂度为 dn^2,时间开销并没有增大。IM_SVM 算法用区间中值衡量区间型数据的相似性,只考虑了区间型数据的内部,丢失了区间大小这一关键信息;IBV_SVM 算法则是用左右边界值衡量区间型数据的相似性,虽然考虑到区间大小,但是内部分布没有涉及;GIK_SVM 算法用区间中值与半宽度衡量区间型数据的相似性,综合了这两方面的信息。因此,结合以上的时间复杂度,GIK_SVM 算法相对更加全面可行。

2 实验结果及分析

2.1 实验环境及实验数据

所有程序均在 Matlab R2014a 平台下实现。实验环境是联想台式电脑,CPU 为 Inter(R) Core (TM) i7-4790,3.60 GHz,内存为 8.00 GB。本文实验采用 4 个人造数据集和 2 个真实数据集。对于 4 个人造数据集,规模均为 12 000,包含 2 个类。两类数据均值分别位于(0.25,0.25)和(0.75,0.75)处,4 组数据的方差分别为(0.1,0.1),(0.2,0.2),(0.5,0.5)和(1.0,1.0),按照高斯分布随机产生区间中值 $\bar{x}_i=(\bar{b}_{i1},\bar{b}_{i2})$,区间半宽度 $\hat{x}_i=(\hat{b}_{i1},\hat{b}_{i2})$ 在[0,0.15]的均匀分布上随机生成,4 个人造数据集的两维特征值都为区间型数据,生成方法为 $([\bar{b}_{i1}-\hat{b}_{i1},\bar{b}_{i1}+\hat{b}_{i1}],[\bar{b}_{i2}-\hat{b}_{i2},\bar{b}_{i2}+\hat{b}_{i2}])$。图 1 为 4 个人造数据集上随机选取 200 个数据的分布图。真实数据集来源于 "Reliable Prognosis" 站点(rp5.ru)提供的气象数据[15],其中一组是哈尔滨-三亚数据集(HS_Ds),该数据集选取了哈尔滨(机场)和三亚(气象站)两个城市自 2006～2015 年以来的统计数据,包括气温与湿度两个指标;另一组太原市-北京数据集(TB_Ds)选取了太原市(机场)和北京两个城市自 2006～2015 年以来的统计数据,包括气温、大气压、湿度、水平能见度和露点温度(空气中的水蒸气变为露珠时的温度)5 个指标。本文实验使用的数据集如表 1 所示。为了保证方法的稳定性,本文的每个实验都是 5 次实验结果的平均值。

表 1 实验使用的数据集

Tab. 1 Datasets used in experiments

数据集	训练样本数	测试样本数	维数
Ds1	9 600	2 400	2
Ds2	9 600	2 400	2
Ds3	9 600	2 400	2
Ds4	9 600	2 400	2
HS_Ds	5 842	1 460	2
TB_Ds	5 842	1 460	5

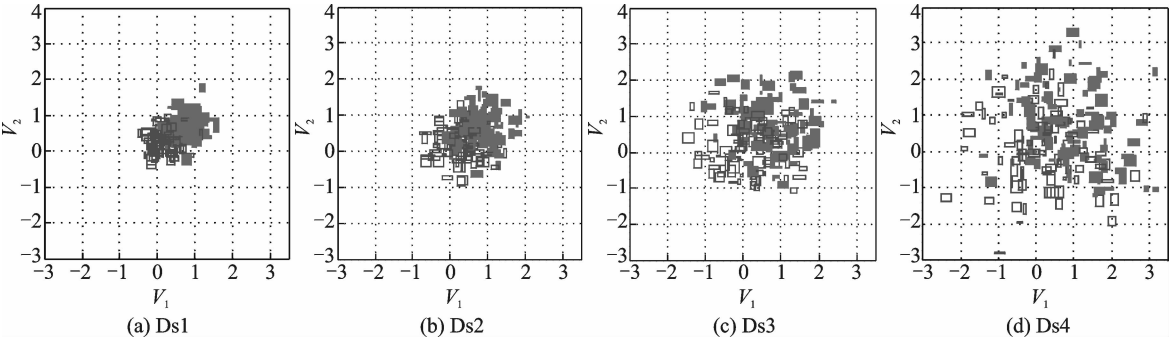


图 1 人造数据集的分布

Fig. 1 Distributions of artificial datasets

2.2 半宽因子对算法的影响

本实验中，TB_Ds 上 $\sigma=0.1$，在其他数据集上 $\sigma=0.25$。为简单起见，本文实验中令 $\alpha_1=\alpha_2=\cdots=\alpha_k=\alpha$，用 α 来表示半宽因子。图 2 为 3 种算法预测准确率随 α 变化的实验结果。由图 2 可以看出 IM_SVM 算法和 IBV_SVM 算法与半宽因子 α 值无关，所以其准确率曲线不发生变化。GIK_SVM 的预测准确率随着半宽因子 α 的调整不断变化，在 Ds1 上，GIK_SVM 整体优于另两种算法，在 $\alpha=0.05$ 处，取到最优准确率；在 Ds2 上，GIK_SVM 介于另两种算法之间，在 $\alpha=0.01$ 处，GIK_SVM 算法的准确率也达到了最优值；在 Ds3 与 Ds4 这两个数据集上，GIK_SVM 浮动较大，Ds3 中，在 $\alpha=0.1$ 与 $\alpha=0.75$ 处，GIK_SVM 算法的准确率优于另两种算法，且当 $\alpha=0.75$ 时，达到最优；在 Ds4 中，当 $\alpha=0.025$ 时，GIK_SVM 虽与 IM_SVM 算法十分接近，但仍取到最优准确率。从这 4 个图中也能看出，除 Ds2 外，本文算法的最优准确率明显高于另外两个算法。从它们的分布来看，Ds1 中两类数据分布紧密，但界限很清晰；Ds3 与 Ds4 中则是混合重叠较多，数据较分散；而 Ds2 中两类数据不仅分布紧密，还混合重叠较多，这也造成了 GIK_SVM 方法的分类准确率不如在其他人造数据集上效果好，且不如 IBV_SVM 方法准确率高。IM_SVM 算法在两个人造数据集上的准确率优于 IBV_SVM 算法，而在另外两个人造数据集上的准确率则不如 IBV_SVM 算法。由于人造数据集的构造方法较简单，使得本文算法与另两种算法在人造数据集上的比较结果相差并不大，因此，本文又在真实数据集上进行了实验。

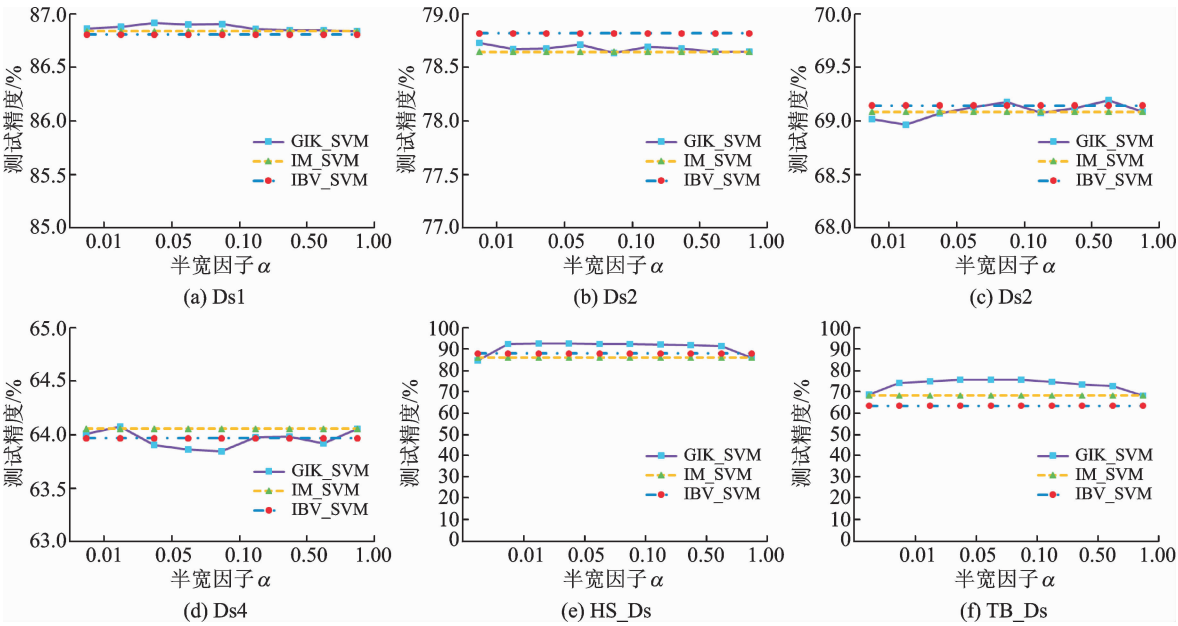


图 2 3 种算法的预测准确率随 α 变化的实验结果

Fig. 2 Experimental results of the prediction accuracy with α for three algorithms

在真实数据集上，除去 α 在两个边界 0 和 1 的取值外，GIK_SVM 算法结果要比另两个算法效果都好。α 为 0 时，算法中只有区间半宽度参与计算，而区间中值没有影响；α 为 1 时，区间半宽度对区间中值不起任何作用，相当于 IM_SVM 算法；当 $\alpha\in(0,1)$ 时，由于区间核中 α 因子的调节，使区间中值与半宽度达到一定程度的平衡，其效果要比只考虑单一条件要好。IM_SVM 算法在 TB_Ds 上的准确率优于 IBV_SVM 算法，而在 HS_Ds 上不如 IBV_SVM 算法的准确率。从以上结果可以看出，本文提出的 GIK_SVM 算法在真实数据集上的结果要好于人造数据集上的实验结果。GIK_SVM 方法可以通过半宽因

子 α 对区间中值与区间半宽度进行折中,寻找一个更适合于区间型数据的衡量指标,以进一步构造高效的算法。另外,从实验中还可以看出,3 种方法的效果都与数据的离散程度有关:离散程度越大,准确率越低;随着数据的离散程度增大,IM_SVM 算法与 IBV_SVM 算法的准确率时好时坏,不太稳定,而 GIK_SVM 算法表现得更为稳定。

2.3 参数 σ 对算法的影响

本实验主要考察参数 σ 对算法的影响。选取图 2 中 GIK_SVM 的最优准确率对应的 α 值作为本实验各数据集的默认 α 值,图 3 为 3 种算法预测准确率随 σ 变化的实验结果。由图 3 可以看出不同的 σ 值对准确率的影响很大。当 $\sigma<2$ 时,在 Ds1 与 2 个真实数据集上,GIK_SVM 预测准确率较高;在其余数据集上,3 种算法的准确率相当。当 $\sigma>2$ 时,3 种算法的准确率都开始降低。在 Ds1 上,GIK_SVM 仍优于另外两种算法;在 HS_Ds 上,GIK_SVM 介于另两种算法之间;在 TB_Ds 上,3 种算法的准确率相当;在其他 3 个数据集上,IBV_SVM 算法准确率快速降低,GIK_SVM 与 IM_SVM 算法准确率大体一致。目前关于高斯核参数 σ 的优化已有很多研究,而本文主要关注高斯区间核度量区间型数据相似性的有效性,所以未对 σ 进行进一步优化,后续实验中,本文选取 $\sigma<2$ 的值。

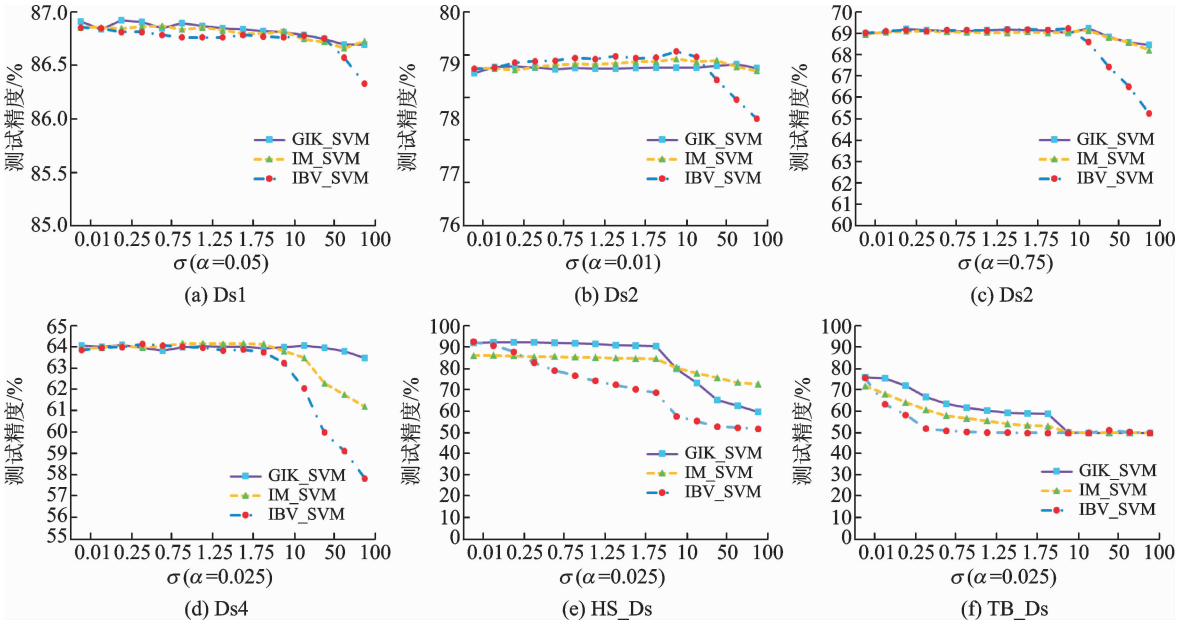


图 3 3 种算法预测准确率随 σ 变化的实验结果

Fig. 3 Experimental results of prediction accuracy with σ for three algorithms

2.4 本算法与决策树模型的比较

为了进一步验证 GIK_SVM 算法的有效性,本文还与 3 种决策树模型进行比较。基于中值半宽的决策树模型(Decision tree based on interval median and boundary value, IMBV_DT)将中值与半宽度分别作为判别属性;基于中值的决策树模型(Decision tree based on interval median, IM_DT)只将中值作为判别属性;基于边界值的决策树模型(Decision tree based on interval boundary value, IBV_DT)将区间的上下边界值作为判别属性。本实验中,TB_Ds 上 $\sigma=0.1$,在其他数据集上 $\sigma=0.25$。图 4 为 GIK_SVM 算法与决策树模型预测准确率随 α 值变化的实验结果。由图 4 可以看出 GIK_SVM 算法的准确率明显高于 3 种决策树模型,其准确率高出 5%~9%左右。在 4 个人造数据集上,GIK_SVM 方法准确

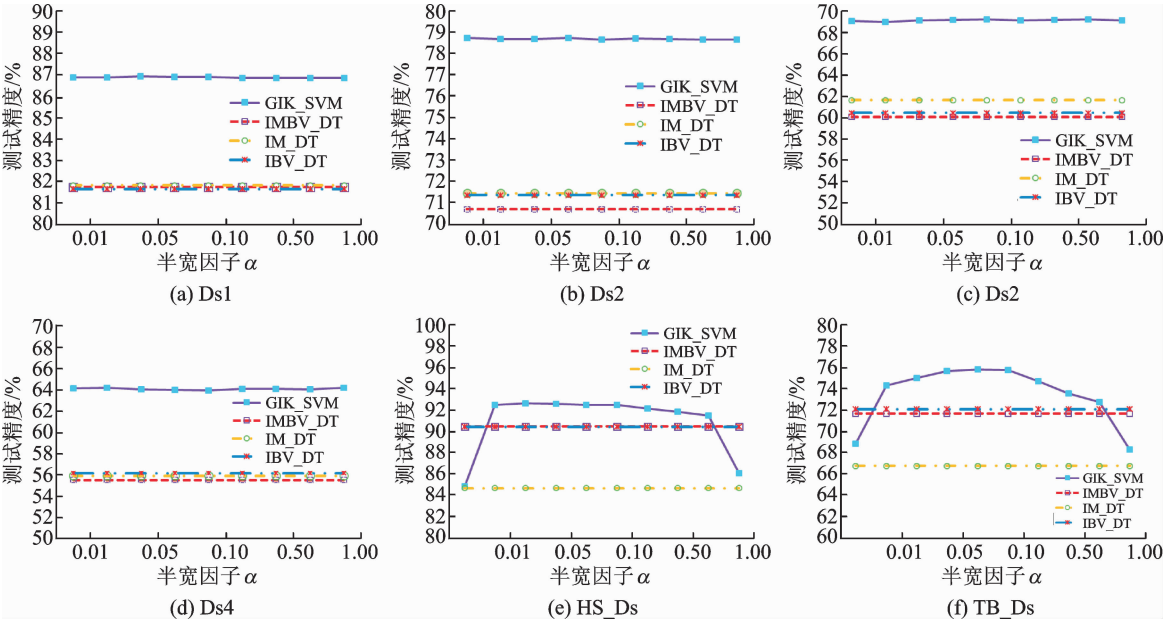


图 4 GIK_SVM 与决策树模型预测准确率随 α 值变化的实验结果

Fig. 4 Experimental results of prediction accuracy with α for GIK_SVM and decision tree models

率曲线都高于其余 3 条曲线，决策树模型的 3 条曲线则比较邻近。在两个真实数据集上，除去 α 取到边界值 0 和 1 时，GIK_SVM 算法的准确率也都明显比决策树模型高，即使在边界值上，GIK_SVM 的准确率仍高于 IM_DT 算法。实验结果最终表明，SVM 模型优于决策树模型。

3 结 论

区间型数据是一类常见然而较为特殊的数据形式，目前关于区间型数据处理的高效分类方法研究还相对较少。本文通过引入半宽因子，很好地折中了区间中值与区间半宽度对区间型数据挖掘的影响，此外构建了高斯区间核，并用高斯区间核 SVM 模型对区间型数据进行分类，提高了分类预测性能。在后续研究中，将考虑针对不同的区间型特征值，选取不同的半宽因子，以取得更好的效果。另外，将探索构造更多的有效区间核，进一步提高处理区间型数据的有效性。

参考文献：

[1] 何清，李宁，罗文娟，等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能，2014，27(4)：327-336.
He Qing，Li Ning，Luo Wenjuan，et al. A survey of machine learning algorithms for big data[J]. Pattem Recognition and Aitificial Intelligence，2014，27(4)：327-336.

[2] 潘志松，唐斯琪，邱俊洋，等. 在线学习算法综述[J]. 数据采集与处理，2016，31(6)：1067-1082.
Pan Zhisong，Tang Siqi，Qiu Junyang，et al. Survey on online learning algorithms[J]. Journal of Data Acquisition and Processing，2016，31(6)：1067-1082.

[3] 胡凯，孟广武，于西昌. 区间值模糊集上的上(下)近似[J]. 模糊系统与数学，2007，21(1)：123-127.
Hu Kai，Meng Guangwu，Yu Xichang. Upper (lower) approximation of an interval-valued fuzzy set[J]. Fuzzy System and Mathematics，2007，21(1)：123-127.

[4] Billard L，Diday E. Regression analysis for interval-valued data[C]// Data Analysis，Classification and Related Methods，Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS'00). Berlin，Heidelberg：Springer-Verlag Press，2000：369-374.

[5] Billard L，Diday E. Symbolic regression analysis[C]// Classification，Clustering and Data Analysis，Proceedings of the

Eighth Conference of the International Federation of Classification Societies (IFCS'02). Berlin, Heidelberg: Springer-Verlag Press, 2002: 281-288.

[6] Cabanes G, Bennani Y, Destenay R. A new topological clustering algorithm for interval data[J]. Pattern Recognition, 2013, 46(11):3030-3039.

[7] Carvalho F D A T D. A fuzzy clustering algorithm for symbolic interval data based on a single adaptive Euclidean distance [C]//Proceedings of the 13th International Conference on Neural Information Processing(ICONIP2006). Berlin, Heidelberg: Springer-Verlag Press, 2006: 1012-1021.

[8] Lima Neto E D A, De Carvalho F D A T. Centre and range method for fitting a linear regression model to symbolic interval data[J]. Computational Statistics and Data Analysis, 2008, 52(3):1500-1515.

[9] Domingues M A. O, Souza R M C R D, Cysneiros F J A. A robust method for linear regression of symbolic interval data[J]. Pattern Recognition Letters, 2010, 31:1991-1996.

[10] 吕泽华, 金海, 袁平鹏, 等. 基于 Gauss 分布函数的区间值数据的模糊聚类算法[J]. 电子学报,2010,38(2):295-300.
Lü Zehua, Jin Hai, Yuan Pingpeng, et al. A fuzzy clustering algorithm for interval-valued data based on gauss distribution functions[J]. Acta Electronica Sinica, 2010, 38(2): 295-300.

[11] 于洋, 张颖, 胡舒涵. 区间型数据聚类的 FCM 新算法[C]//中国通信学会第六届学术年会论文集(下). 北京:中国通信学会,2009: 249-253.
Yu Yang, Zhang Ying, Hu Shuhan. The new FCM algorithm of interval data clustering[C]//Proceedings of the 6th Academic Annual Conference of China Communication Association. Beijing: China Institute of Communications Press, 2009: 249-253.

[12] 陈建凯, 王鑫, 何强, 等. 区间值属性的单调决策树算法[J]. 模式识别与人工智能, 2016, 29(1): 47-53.
Chen Jiankai, Wang Xin, He Qiang, et al. Interval-valued attributes based monotonic decision tree algorithm[J]. Pattem Recognition and Aitificial Intelligence, 2016, 29(1): 47-53.

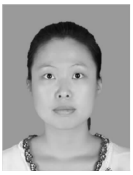
[13] Vapnik V. Statistical learning theory[M]. New York: Springer-Verlag Press, 1998:493-520.

[14] Cortes C, Vapnik V. Support vector networks[J]. Machine Learning, 1995(20): 273-297.

[15] 中国 6603 个居民点天气[EB/OL]. http://rp5.ru/中国天气_, 2016-04.

作者简介:

王文剑(1968-),女,博士,教授,研究方向:神经网络,支持向量机,计算智能和数据挖掘, E-mail:wjwang@sxu.edu.cn。



祁晓博(1992-),女,硕士研究生,研究方向:机器学习和数据挖掘。

郭虎升(1986-),男,博士,副教授,研究方向:支持向量机,机器学习和数据挖掘。