山西大学学报(自然科学版)47(1):18-29,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023138

基于模糊隶属度函数的SVM样本约简算法

张代俐,汪廷华*,朱兴淋
(赣南师范大学 数学与计算机科学学院,江西 赣州 341000)

摘 要:支持向量机(Support vector machine,SVM)具有良好的学习泛化性能,但其学习效率随着训练样本数量的增加而显著降低,对于大规模训练集,采用标准优化方法的传统SVM面临着内存需求过大、执行速度慢等问题。为了缓解这个问题,由于不同的数据点对决策平面的贡献程度不同,本文通过模糊隶属度函数计算每个样本的隶属度,利用模糊隶属度评估每个样本的重要程度,从而将隶属度值低的样本进行约简。基于三种不同的模糊隶属度函数,分别提出了基于类中心距离、核目标对齐和中心核对齐模糊隶属度函数的SVM样本约简算法。在UCI(University of California, lrvine)和kaggle数据集上与传统的SVM和最近提出的基于牛顿法稀疏化SVM(Newton-type Sparse SVM, NSSVM)进行了大量的对比实验,实验结果验证了所提出的基于模糊隶属度函数的SVM样本约简算法在准确率、*F*-度量和Hinge损失这几个分类性能指标方面的优势。例如,基于中心核对齐模糊隶属度的SVM约简算法在diabetes数据集上取得了最高的准确率、*F*-度量和最小的Hinge损失。与SVM相比,准确率和*F*-度量分别提高了13.71%和9.55%,Hinge损失降低了3.28%;与NSSVM相比,准确率和*F*-度量分别提高了24.54%和9.38%,Hinge损失降低了21.54%。

关键词:机器学习;支持向量机;样本约简;模糊隶属度函数
中图分类号:TP181 **文献标志码:**A **文章编号:**0253-2395(2024)01-0018-12

SVM Sample Reduction Algorithm Based on Fuzzy Membership Functions

ZHANG Daili, WANG Tinghua*, ZHU Xinglin
(*School of Mathematics and Computer Science, Gannan Normal University, Ganzhou* 341000*, China*)

Abstract: Support vector machine (SVM) has good learning generalization performance. However, the learning efficiency of SVM decreases significantly with the increase of the number of training samples. For large-scale training sets, the traditional SVM with standard optimization methods confronts problems such as excessive memory requirements and slow training speed. In order to alleviate this problem, due to the different contribution of different data points to the decision plane, in this paper, we calculate the fuzzy membership of each sample through the fuzzy membership function, and use the fuzzy membership to evaluate the importance of each sample, so as to delete the samples with low memberships. Based on three different fuzzy membership functions, SVM sample reduction algorithms based on class center distance, kernel target alignment and centered kernel alignment fuzzy membership functions are proposed, respectively. Comprehensive comparative experiments are performed on UCI (University of California, lrvine) and kaggle data sets with the traditional SVM and the proposed Newton-type Sparse SVM (NSSVM). The experimental results validate the advantages of the proposed SVM sample reduction algorithms based on fuzzy membership functions in terms of Accuracy, *F*-measure and Hinge loss measures. For example, the algorithm based on the centered kernel alignment fuzzy membership function

收稿日期:2023-06-05;**接受日期:**2023-08-09
基金项目:国家自然科学基金(61966002);江西省研究生创新专项资金(YC2022-s944)
作者简介:张代俐(1996-),女,江西赣州人,硕士研究生,研究方向为机器学习与数据挖掘。E-mail:17746670679@163.com
* **通信作者:**汪廷华(WANG Tinghua),E-mail:wthpku@163.com
引文格式:张代俐,汪廷华,朱兴淋.基于模糊隶属度函数的SVM样本约简算法[J].山西大学学报(自然科学版),2024,47(1):18-29. DOI:10.13451/j.sxu.ns.2023138

achieves the highest Accuracy, *F*-measure, and the smallest Hinge loss on the diabetes data set. Compared with the SVM, the Accuracy and *F*-measure are increased by 13.71% and 9.55%, respectively, and the Hinge loss is reduced by 3.28%. Compared with the NSSVM, the accuracy and *F*-measure are increased by 24.54% and 9.38%, respectively, and the Hinge loss is reduced by 21.54%.
Key words: machine learning; support vector machine (SVM); sample reduction; fuzzy membership function

0 引言

支持向量机(Support vector machine,SVM)是由Vapnik等在20世纪90年代提出来的一种有监督的学习方法。SVM以统计学习理论为基础,基于VC维(Vapnik-Chervonenkis dimension)理论和结构风险最小原理,很大程度上克服了局部极小和维数灾难问题[1-2]。由于SVM良好的理论基础和泛化性能,因此被广泛应用于计算机视觉[3-4]、文本分类[5-6]、故障诊断[7-8]、面部识别[9-10]等领域。SVM通过计算最近训练样本到分类超平面的最大化间隔来构造分类超平面,而分类超平面仅由支持向量(Support vector,SV)决定。SVM适用于线性可分和非线性可分的数据分类。从几何的角度上看,对于一个线性可分的数据集,为每个类标签的数据构造凸包(Convex hull,CH),并选择凸包中最近的数据点,最优超平面是正交平分位于最近点之间的平面[11]。针对现实任务中的非线性可分的数据集,SVM引入核技巧[12-13],通过一个非线性变换将原样本空间映射到高维特征空间,然后在特征空间中实施线性学习算法,构造最优分类超平面,并运用核函数计算样本点在特征空间中的内积,降低计算的复杂性。虽然核技巧提高了SVM分类器对非线性可分数据的分类精度,但由于映射数据点需要大量的计算,显著增加了训练时间和计算开销。

传统的SVM的本质是求解凸二次规划问题。二次规划问题的标准求解方法,需要对大小为$t \times t$(t为训练样本数)的Gram矩阵进行存储和优化计算,矩阵存储所占用的内存空间随样本个数增加而平方增长[14]。另外,凸二次规划的优化计算时间也随训练样本的增多而快速增加,其求解算法的时间复杂度达到了$O(t^3)$[15]。因此,SVM分类器的效率随着数据点数量的增加而降低,对于大规模训练集情况,采用标准优化方法的传统SVM将面临着内存需求过大、执行速度慢,有时甚至无法执行的问题[16]。尽管存在旨在加速SVM训练的优化技术,其中包括chunking[17]、并行[18-19]和修改工作集[20]的方法,但它们通常在优化过程中引入了额外的内存负担。

Zhou[21]考虑了稀疏约束核支持向量机优化,以控制支持向量的数量。基于建立的与平稳方程相关的最优性条件,开发了一种牛顿型方法来处理稀疏约束优化的基于牛顿法稀疏化SVM(Newton-type Sparse SVM, NSSVM)算法。若起点被选择为接近静止点的局部区域,则该方法具有一定的收敛特性,从而导致超高的计算速度。但该算法的收敛性跟起点的选择有很大的关系,具有不稳定性。由SVM的分类原理可以知道,一般情况下,并不是所有的训练样本都参与了决策判别,支撑分类超平面的所谓支持向量仅仅只是全部训练集中的一小部分关键点。如果预先对训练样本有所选择,不仅可以减少存储空间占用,而且能够节省计算时间。针对样本集的约简,Lee等[22]提出的简化支持向量机(Reduced SVM,RSVM)通过随机选择训练数据的子集用于训练获得分离超平面。实验表明,尽管训练数据集大量减少了,但RSVM在测试集上的结果甚至比整个数据集获得的结果更好,这可能是由于减少了数据的过拟合。但是由于RSVM算法对初始样本集的选择和样本分布很敏感,对样本分布不均匀的训练集进行随机抽取效果不是很理想。为了解决这一问题,Goodrich等[23]结合数据的几何分布,考虑边界区域上的样本更有可能是支持向量的特性,提出简化凸壳(Reduced Convex Hull,RCH)算法,从而能够快速去除距离分类边界较远的样本,除此之外,该算法还能通过改变上界因子的大小减少异类凸包的重叠,从而将线性不可分转化为线性可分情形。尽管RCH算法综合考虑样本的分布,但其容易受噪声和离群点的影响,在不平衡数据集上适用性

不高。Bang等[24]考虑将其与聚类算法相结合，提出基于加权k均值聚类的SVM算法(Weighted k-means clustering based SVM)算法，该算法通过使用每个簇中的数据点数量作为权重，对误分类样本施加不同的惩罚。实验证明，该算法有效应用于处理不平衡数据集。同时，Cervantes等[25]提出的算法采用了与序列最小化算法(Sequential minimal optimization，SMO)类似的思想，通过将训练集划分成小簇来处理大规模数据集，该算法应用模糊聚类算法去除了聚类中心远离最优超平面的簇，保留距离分类超平面距离较近的簇。结果表明，该算法的训练时间明显小于传统的SVM，而且获得的支持向量数量相似。目前，基于聚类的SVM样本约简算法对大规模训练集的约简做出了巨大的贡献，模糊聚类可以有效地处理噪声和离群点对分类器性能的影响。受模糊支持向量机(Fuzzy SVM，FSVM)[26]启发，考虑使用模糊隶属度对样本的重要程度进行评估，从而将隶属度值较低的样本约简，既可以减少噪声，又能减少训练集的大小。

本文的目标是提出一种基于模糊隶属度函数的SVM样本约简算法，用于SVM分类。模糊隶属度函数用于计算样本的模糊隶属度值，从而评估每个样本的重要程度。文献[27]通过样本到类中心距离的大小来确定样本的隶属度大小。Lin和Wang[28]提出了一种使用核目标对齐(Kernel target alignment，KTA)[29]的模糊隶属度计算方法，其中使用从KTA导出的启发式函数来计算数据点和该点的标签之间的对齐。Wang等[30]使用中心核对齐(Centered kernel alignment，CKA)代替KTA方法来计算数据点与其关联标签之间的依赖关系，并为FSVM开发了一种新的模糊隶属度计算方法。

本文的主要创新和贡献是将FSVM中的模糊隶属度函数应用于SVM样本约简中，提出了基于类中心距离、KTA和CKA三种模糊隶属度函数的SVM样本约简算法。所提出的算法利用模糊隶属度评估每个训练样本的重要程度，将模糊隶属度值低的训练样本约简，从而减少算法的存储空间和计算时间。

1 支持向量机

假设训练样本集合$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_t,y_t)\}$由$t$个训练样本$\boldsymbol{x}_i$组成，其中$\boldsymbol{x}_i\in\mathrm{R}^d$，$i\in1,\cdots,t$，以二分类问题为例，$\boldsymbol{x}_i$对应的类别标签为$y_i\in\{-1,+1\}$，$y_i=-1$为负类样本，$y_i=+1$为正类样本。SVM分类器的基本思想是基于训练集$\boldsymbol{T}$在样本空间或特征空间中找到一个最优超平面$\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b=0$，使得该超平面距离正负两类样本的间隔最大。其中，$\boldsymbol{w}$为权重向量，$\phi(\boldsymbol{x})$为$\boldsymbol{x}$在特征空间上的映射，$b$为偏置。最优超平面可通过求解以下的优化问题得到：

$$\begin{aligned}&\min_{\boldsymbol{w},b,\xi_i}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{t}\xi_i,\\&\text{s.t.}\,y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\geqslant1-\xi_i,\\&\xi_i\geqslant0,i=1,2,\cdots,t,\end{aligned}\tag{1}$$

其中，ξ_i为松弛变量，C是控制边界和松弛惩罚之间权衡的参数。为解决上述SVM优化问题，引入拉格朗日对偶将上述约束问题转化为以下无约束问题：

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{t}\alpha_i-\frac{1}{2}\sum_{i=1}^{t}\sum_{j=1}^{t}\alpha_i\alpha_jy_iy_j\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j),\\&\text{s.t.}\sum_{i=1}^{t}\alpha_iy_i=0,\\&0\leqslant\alpha_i\leqslant C,\ i=1,2,\cdots,t。\end{aligned}\tag{2}$$

为避免在高维特征空间计算困难，SVM引入核技巧，通过核函数$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$隐式定义特征映射$\phi$，而不必显式计算出$\phi$的具体值。SVM的决策函数可以被定义为：

$$f(\boldsymbol{x})=\mathrm{sign}\left(\sum_{i=1}^{t}\alpha_iy_i\kappa(\boldsymbol{x},\boldsymbol{x}_i)+b\right)。\tag{3}$$

由式(1)和(2)可知，SVM求解传统的二次规划问题，但对于大规模的真实数据集，将变得不可行。因此，减少SVM训练集大小的算法被认为是解决从大规模数据集中学习SVM问题的直接有效措施。

2 SVM样本约简算法设计

本节先介绍三种常用的模糊隶属度函数，然后提出基于模糊隶属度函数的SVM样本约简算法。

2.1 基于类中心距离的模糊隶属度函数

基于类中心距离的模糊隶属度函数的基本思想是通过样本到类中心距离的大小来确定样本的隶属度大小,若样本到类中心的距离越小,样本的隶属度值越大,反之,则越小。假设C^+和C^-分别为正类和负类样本的集合,T_+和T_-分别为正负类样本的类中心,d_{i+}和d_{i-}分别表示正负类样本到其所在类中心的距离,r_+和r_-分别为正负类样本到其类中心的最远距离,t_+和t_-分别为正负类样本个数。则有:

$$\begin{cases} C^+=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in T \text{ and } y_i=+1\}, \\ C^-=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in T \text{ and } y_i=-1\}。 \end{cases} \tag{4}$$

$$\begin{cases} T_+=\dfrac{1}{t_+}\sum\limits_{i=1}^{t_+}\boldsymbol{x}_i,\boldsymbol{x}_i\in C^+, \\ T_-=\dfrac{1}{t_-}\sum\limits_{i=1}^{t_-}\boldsymbol{x}_i,\boldsymbol{x}_i\in C^-, \end{cases} \tag{5}$$

$$\begin{cases} d_{i+}=\|\boldsymbol{x}_i-T_+\|,\boldsymbol{x}_i\in C^+, \\ d_{i-}=\|\boldsymbol{x}_i-T_-\|,\boldsymbol{x}_i\in C^-, \end{cases} \tag{6}$$

$$\begin{cases} r_+=\max d_{i+}, \\ r_-=\max d_{i-}。 \end{cases} \tag{7}$$

类似地,对于高维特征空间,通过引入核技巧,将样本点$\boldsymbol{x}_i$映射为$\phi(\boldsymbol{x}_i)$,则正负类的样本中心变成相应的$\phi(T_+)$和$\phi(T_-)$,则有:

$$\begin{cases} \phi(T_+)=\dfrac{1}{t_+}\sum\limits_{i=1}^{t_+}\phi(\boldsymbol{x}_i),\boldsymbol{x}_i\in C^+, \\ \phi(T_-)=\dfrac{1}{t_-}\sum\limits_{i=1}^{t_-}\phi(\boldsymbol{x}_i),\boldsymbol{x}_i\in C^-。 \end{cases} \tag{8}$$

$$\begin{cases} d_{i+}=\|\phi(\boldsymbol{x}_i)-\phi(T_+)\|,\boldsymbol{x}_i\in C^+, \\ d_{i-}=\|\phi(\boldsymbol{x}_i)-\phi(T_-)\|,\boldsymbol{x}_i\in C^-。 \end{cases} \tag{9}$$

由于$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$,故有

$$\begin{aligned} d_{i+}= \\ \sqrt{\|\phi(\boldsymbol{x}_i)-\phi(T_i)\|^2}= \\ \sqrt{\phi(\boldsymbol{x}_i)^2-2\phi(\boldsymbol{x}_i)\cdot\phi(T_i)+\phi(T_i)^2}= \\ \sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)-2\kappa(\boldsymbol{x}_i,T_i)+\kappa(T_i,T_i)}。 \end{aligned} \tag{10}$$

基于类中心距离的模糊隶属度函数可以表示为s_i:

$$s_i=\begin{cases} 1-\dfrac{d_{i+}}{r_{i+}+\delta},\boldsymbol{x}_i\in C^+, \\ 1-\dfrac{d_{i-}}{r_{i-}+\delta},\boldsymbol{x}_i\in C^-。 \end{cases} \tag{11}$$

其中,δ为事先给定的一个极小的正数。

2.2 基于核对齐的模糊隶属度函数

核目标对齐是由Cristianini等[29]首次提出。核矩阵(Gram matrix)通过核函数κ定义为$\boldsymbol{K}_{i,j}=\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$。给定两个核函数$\kappa_1$和$\kappa_2$,核目标对齐定义为

$$K_{\mathrm{KTA}}(\boldsymbol{K}_1,\boldsymbol{K}_2)=\frac{\langle\boldsymbol{K}_1,\boldsymbol{K}_2\rangle_F}{\sqrt{\langle\boldsymbol{K}_1,\boldsymbol{K}_1\rangle\langle\boldsymbol{K}_2,\boldsymbol{K}_2\rangle}_F}。 \tag{12}$$

其中$\langle\boldsymbol{K}_1,\boldsymbol{K}_2\rangle_F=\sum\limits_{i=1}^{t}\sum\limits_{j=1}^{t}\kappa_1(\boldsymbol{x}_i,\boldsymbol{x}_j)\kappa_2(\boldsymbol{x}_i,\boldsymbol{x}_j)$。文献[28]引入启发式函数$h(\boldsymbol{x})$,同时在训练数据点的过程中引入置信因子$h_{\mathrm{C}}$和无用因子$h_{\mathrm{T}}$,并通过使用这两个因子和映射函数,从启发式函数自动生成训练数据点的模糊隶属度。假设$p_x(\boldsymbol{x})$为非噪声数据点$\boldsymbol{x}$的概率密度函数,$p_x(\boldsymbol{x})$可被定义为:

$$p_x(\boldsymbol{x})=\begin{cases} 1, \text{ if } h(\boldsymbol{x}_i)>h_{\mathrm{C}}, \\ \rho, \text{ if } h(\boldsymbol{x}_i)<h_{\mathrm{T}}, \\ \rho+(1-\rho)\left(\dfrac{h(\boldsymbol{x})-h_{\mathrm{T}}}{h_{\mathrm{C}}-h_{\mathrm{T}}}\right)^{\beta}, \text{ otherwise}。 \end{cases} \tag{13}$$

其中$h(\boldsymbol{x})=\sum\limits_{i=1}^{t}\sum\limits_{j=1}^{t}y_iy_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$,$\rho$为隶属度的下限,$\beta$为多项式的系数。通过式(13)计算得到数据点的概率密度值作为对样本的重要程度的估计,再根据式(14)对其进行归一化得到样本的隶属度:

$$s_j=\frac{d_j-\min\{d_i\}_{i=1}^{t}}{\max\{d_i\}_{i=1}^{t}-\min\{d_i\}_{i=1}^{t}}, \tag{14}$$

其中,d_j为使用KTA模糊隶属度函数计算得到的模糊隶属度值,$\min\{d_i\}_{i=1}^{t}$和$\max\{d_i\}_{i=1}^{t}$分别为训练数据集的最小和最大的模糊隶属度值。

2.3 基于中心核对齐的模糊隶属度函数

文献[30]使用CKA代替KTA方法来计算数据点与其关联标签之间的依赖关系,并为FSVM开发了一种新的模糊隶属度计算方法。该方法使用$\overline{\boldsymbol{K}}=\mathrm{HKH}$定义中心核矩阵,$\boldsymbol{H}=\boldsymbol{I}-\dfrac{\boldsymbol{e}\boldsymbol{e}^{\mathrm{T}}}{t}$,其中$\boldsymbol{I}\in R^{t\times t}$为单位矩阵,$\boldsymbol{e}=(1,\cdots,1)^{\mathrm{T}}\in R^t$。中心核对齐可定义为:

$$C_{\mathrm{CKA}}(\boldsymbol{K}_1,\boldsymbol{K}_2)=\frac{\langle \overline{\boldsymbol{K}}_1,\overline{\boldsymbol{K}}_2\rangle_F}{\sqrt{\langle \overline{\boldsymbol{K}}_1,\overline{\boldsymbol{K}}_1\rangle_F\langle \overline{\boldsymbol{K}}_2,\overline{\boldsymbol{K}}_2\rangle_F}}。\tag{15}$$

对于二分类问题,给定核矩阵$\boldsymbol{L}\in R^{t\times t}$,$\boldsymbol{L}_{i,j}=h(y_i,y_j)$,其中$h(y_i,y_j)$定义如下:

$$h(y_i,y_j)=\begin{cases}+1, & y_i=y_j,\\ -1, & y_i\neq y_j。\end{cases}\tag{16}$$

式(16)将同类样本的相似性设为1,反之,则设为−1。假设$\boldsymbol{y}=(y_1,\cdots,y_t)^{\mathrm{T}}$,则CKA可表示为:

$$C_{\mathrm{CKA}}(\boldsymbol{K},\boldsymbol{L})=\frac{\langle \overline{\boldsymbol{K}},\overline{\boldsymbol{L}}\rangle_F}{\sqrt{\langle \overline{\boldsymbol{K}},\overline{\boldsymbol{K}}\rangle_F\langle \overline{\boldsymbol{L}},\overline{\boldsymbol{L}}\rangle_F}}=\frac{\langle \overline{\boldsymbol{K}},\boldsymbol{L}\rangle_F}{\sqrt{\langle \overline{\boldsymbol{K}},\boldsymbol{K}\rangle_F\langle \overline{\boldsymbol{L}},\boldsymbol{L}\rangle_F}}=\frac{\langle \overline{\boldsymbol{K}},\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}\rangle_F}{\sqrt{\langle \overline{\boldsymbol{K}},\boldsymbol{K}\rangle_F\langle \overline{\boldsymbol{L}},\boldsymbol{L}\rangle_F}}=\frac{\sum_{i=1}^{t}\sum_{i=1}^{t}y_iy_j\overline{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sqrt{\langle \overline{\boldsymbol{K}},\boldsymbol{K}\rangle_F\langle \overline{\boldsymbol{L}},\boldsymbol{L}\rangle_F}}=\frac{1}{\sqrt{\langle \overline{\boldsymbol{K}},\boldsymbol{K}\rangle_F\langle \overline{\boldsymbol{L}},\boldsymbol{L}\rangle_F}}\left[\sum_{y_i=y_j}\overline{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)-\sum_{y_i\neq y_j}\overline{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)\right]。\tag{17}$$

其中,$\overline{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)$为中心核函数,并且$\overline{\kappa}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\overline{\boldsymbol{K}}_{i,j}$。对于给定的一个输入样本$(\boldsymbol{x}_p,y_p)$,其中$1\leqslant p\leqslant t$,由于系数$\frac{1}{\sqrt{\langle \overline{\boldsymbol{K}},\boldsymbol{K}\rangle_F\langle \overline{\boldsymbol{L}},\boldsymbol{L}\rangle_F}}$是一个常数,当核函数$\kappa$给定后,每个样本的CKA($I$-$C_{\mathrm{CKA}}$)可表示成:

$$I-C_{\mathrm{CKA}}(\boldsymbol{K},\boldsymbol{L},\boldsymbol{x}_p)=\sum_{y_p=y_i}\overline{\kappa}(\boldsymbol{x}_p,\boldsymbol{x}_i)-\sum_{y_p\neq y_i}\overline{\kappa}(\boldsymbol{x}_p,\boldsymbol{x}_i)。\tag{18}$$

通过计算上式计算每个样本的I-C_{CKA}值,再使用式(14)进行归一化,得到样本的隶属度值。

2.4 SVM样本约简算法

根据上述模糊隶属度函数,设计的SVM样本约简算法如表1所示。其中步骤4中模糊隶属度阈值的设定需根据实验需求而定,本实验阈值是按照实验样本数的10%设定的,即约简10%的样本,保留90%的样本。算法的流程如图1所示。

表1 算法伪代码

Table 1 The pseudo code of algorithm

基于模糊隶属度函数的SVM样本约简算法
输入:训练集T,核函数κ,正则化参数C,阈值θ
输出:SVM分类器$f(\boldsymbol{x})$
1:最优化正则化参数C和相应的核参数
2:数据预处理
3:根据上述模糊隶属度函数,计算每个训练样本的模糊隶属度值
4:设定模糊隶属度的阈值θ,将样本模糊隶属度低于θ的样本约简
5:按照7:3的比例划分训练集和测试集
6:将训练集用于SVM训练,并将此SVM分类器用于测试集上测试

3 实验

3.1 实验设置

在本节中,使用实验环境为Windows 10,内存8 GB,MATLAB(R2018a)和PyCharm Community Edition 2021.3,对二分类问题进行了大量的实验,以评估所提出算法的有效性。SVM常用的核函数包括线性核、多项式核和径向基核函数(RBF核)。在下文中,考虑使用RBF核

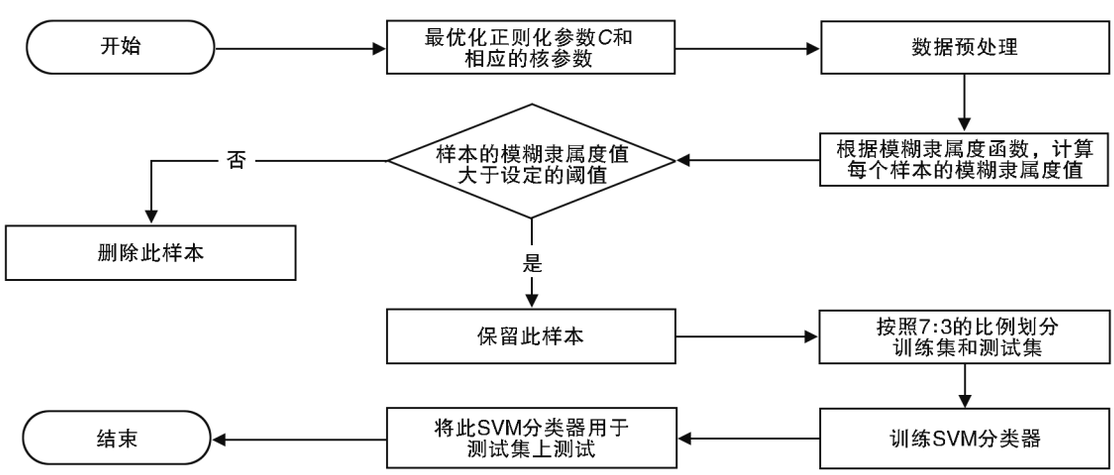


图1 算法流程图

Fig. 1 The flow chart of algorithm

$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2\right)$,其中$\sigma$为宽度参数。根据以上三种不同的模糊隶属度函数,分别提出了基于类中心距离模糊隶属度函数、核目标对齐模糊隶属度函数和中心核对齐模糊隶属度函数的SVM样本约简算法。为书写简洁,我们使用Al-1、Al-2和Al-3分别表示基于类中心距离模糊隶属度函数的SVM样本约简算法、基于KTA模糊隶属度函数的SVM样本约简算法和基于CKA模糊隶属度函数的SVM样本约简算法。同时,将所提出的三种算法与传统的SVM算法和最近提出的NSSVM样本约简算法进行比较:

SVM:传统SVM分类器[1];

NSSVM:通过建立与平稳方程相关的最优性条件,利用牛顿法来处理稀疏约束优化的SVM优化方法[21]。

对于传统的SVM分类器,通过对训练集进行十折交叉验证来确定核参数σ和正则化参数C[31],在本实验中,我们在两个维度上进行网格搜索,即$C=\{2^{-5},2^{-3},\cdots,2^{15}\}$和$\gamma=\{2^{-15},2^{-14},\cdots,2^{3}\}$。其中$\gamma=1/\sigma^2$,以获得最佳参数$C$和$\gamma$。

将提出的基于模糊隶属度函数的三种算法与其他两种算法在准确率(Accuracy)、F-度量(F-measure)和Hinge损失(Hinge-loss)方面进行了分类性能的比较。我们知道,对于给定类别的混淆矩阵(如表2所示),其中的真正例(TP)、假正例(FP)、真反例(TN)和假反例(FN)分别表示正确分类的正类样本数量、错误分类的正类样本数量、正确分类的负类样本数量和错误分类的负类样本数量。分类器精度(Precision)是正确分类的样本数量与分配给该类别的样本总数的比率。召回率(Recall)是正确分类的样本数量与属于该类别的样本总数的比率。F-度量是精度和召回率的调和平均值,其中考虑了精度和召回率,以避免出现高精度和低召回率的情况,反之亦然。准确率是正确分类的样本数量与所有类别的样本总数的比率,这是机器学习和数据挖掘社区中最常用的衡量标准。Hinge损失是SVM常用的一种替代损失函数。形式上,这些准则的定义如下:

$$\text{Precision}=\frac{TP}{TP+FP},\tag{19}$$

表2 混淆矩阵

Table 2 Confusion matrix

	预测正例	预测反例
正例	*TP*	*FN*
反例	*FP*	*TN*

$$\text{Recall}=\frac{TP}{TP+FN},\tag{20}$$

$$F\text{-Measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}},\tag{21}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN},\tag{22}$$

$$\text{Hinge-loss}=\max\left(0,1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\right)。\tag{23}$$

从UCI(University of California, Irvine)(https://archive.ics.uci.edu/ml)数据库中选择了5个二分类数据集数和从kaggle(https://www.kaggle.com/datasets)数据库中选择了1个二分类数据集。表3提供了这些数据集的统计数据,显示了每个数据集的名称,样本数量(正类样本和负类样本的数量在括号中给出),特征数量以及数据来源。

表3 选用的数据集基本信息

Table 3 Basic information of the selected data sets

数据集	样本数量	特征数量	数据来源
wilt	4 839(4 578/261)	5	UCI
twonorm	7 400(3 703/3 697)	20	UCI
titanic	2 201(1 490/711)	3	kaggle
mammographic	961(516/445)	5	UCI
diabetes	768(500/268)	8	UCI
blood	748(570/178)	5	UCI

我们对数据集进行预处理,首先将训练样本的缺失值进行补全,这里使用平均值代替缺失值,然后进行特征归一化。再根据上述模糊隶属度函数对每个样本进行隶属度计算,根据设定的阈值去除隶属度值较低的样本,再将剩余的样本通过分层抽样划分成训练集和测试集(其中数据集的生成遵循类先验概率):70%的训练集和30%的测试集。最后将训练集用于SVM训练,得到的SVM分类器再用于测试集上测试。

3.2 实验结果与分析

为了获得稳定的结果,我们独立地划分每个数据集,然后分别运行每个算法20次。表4记录了每种算法的平均分类结果,粗体数字表示这些方法在每个数据集上的最佳性能。从表

表4 不同方法在各数据集上的分类性能比较(均值±标准差)

Table 4 Comparison of classification performance among different methods on each data set (mean ± standard deviation)

数据集	结果	SVM	NSSVM	All-1	All-2	All-3
wilt	Accuracy	0.975 7±0.002 9	0.945 6±0.005 1	0.986 9±0.002 7	0.990 2±0.002 1	0.990 9±0.001 7
	*F*1-measure	0.987 2±0.001 5	0.970 0±0.002 7	0.993 2±0.001 4	0.994 6±0.001 4	0.994 8±0.001 5
	Hingel-loss	0.639 5±0.019 8	0.462 6±0.067 4	0.601 3±0.142 9	0.427 3±0.080 7	0.271 5±0.070 8
twonorm	Accuracy	0.972 5±0.002 5	0.970 1±0.003 9	0.982 4±0.002 8	0.980 5±0.002 5	0.976 7±0.002 2
	*F*1-measure	0.972 6±0.002 4	0.969 5±0.004 0	0.982 5±0.002 8	0.980 5±0.002 5	0.976 8±0.002 3
	Hingel-loss	0.544 4±0.012 9	0.539 9±0.014 4	0.506 0±0.005 8	0.505 4±0.004 6	0.504 8±0.004 4
titanic	Accuracy	0.789 2±0.006 7	0.683 0±0.028 7	0.827 8±0.011 5	0.825 1±0.010 5	0.766 6±0.007 8
	*F*1-measure	0.863 5±0.004 0	0.790 0±0.024 1	0.886 4±0.006 9	0.884 6±0.006 4	0.852 1±0.004 0
	Hingel-loss	0.585 0±0.029 3	0.381 9±0.063 7	0.250 8±0.006 9	0.267 5±0.004 3	0.324 4±0.004 5
mammographic	Accuracy	0.768 9±0.018 1	0.783 9±0.024 2	0.856 5±0.015 9	0.888 0±0.022 3	0.877 1±0.019 1
	*F*1-measure	0.791 9±0.017 1	0.780 0±0.027 6	0.859 1±0.017 5	0.890 8±0.022 6	0.879 3±0.019 5
	Hingel-loss	0.456 1±0.007 7	0.457 0±0.026 5	0.443 4±0.018 8	0.433 1±0.031 1	0.441 1±0.034 6
diabetes	Accuracy	0.705 2±0.027 1	0.643 9±0.017 6	0.773 3±0.017 2	0.796 1±0.016 7	0.801 9±0.019 2
	*F*1-measure	0.778 8±0.021 3	0.780 0±0.012 9	0.835 8±0.012 2	0.851 8±0.013 5	0.853 2±0.013 8
	Hinge-loss	0.600 9±0.007 5	0.707 1±0.035 3	0.589 1±0.008 2	0.583 2±0.019 1	0.581 8±0.011 3
blood	Accuracy	0.731 4±0.016 2	0.741 1±0.015 7	0.805 2±0.014 1	0.762 9±0.008 8	0.764 9±0.004 1
	*F*1-measure	0.836 0±0.011 2	0.850 0±0.011 9	0.880 6±0.009 4	0.863 8±0.005 8	0.865 3±0.002 4
	Hingel-loss	0.493 3±0.023 1	0.485 7±0.055 2	0.442 7±0.070 2	0.529 4±0.049 2	0.508 8±0.048 7

表5 不同算法在各数据集上的显著性检验

Table 5 Significance test of different algorithms on each data set

Data set	Winl-tiel-loss (Wl-Tl-L)	Al-1 vs. SVM	All-1 vs. NSSVM	All-2 vs. SVM	All-2 vs. NSSVM	All-3 vs. SVM	All-3 vs. NSSVM
wilt	Accuracy	W	W	W	W	W	W
	F-measure	W	W	W	W	W	W
	Hingel-loss	T	W	W	T	W	W
twonorm	Accuracy	W	W	W	W	W	W
	*F*1-measure	W	W	W	W	W	W
	Hingel-loss	W	W	W	W	W	W
titanic	Accuracy	W	W	W	W	L	W
	*F*1-measure	W	W	W	W	L	W
	Hingel-loss	W	W	W	W	W	W
mammographic	Accuracy	W	W	W	W	W	W
	*F*1-measure	W	W	W	W	W	W
	Hingel-loss	W	T	W	W	T	T
diabetes	Accuracy	W	W	W	W	W	W
	*F*1-measure	W	W	W	W	W	W
	Hinge-loss	W	W	W	W	W	W
blood	Accuracy	W	W	W	W	W	W
	*F*1-measure	W	W	W	W	W	W
	Hingel-loss	W	W	L	L	T	T

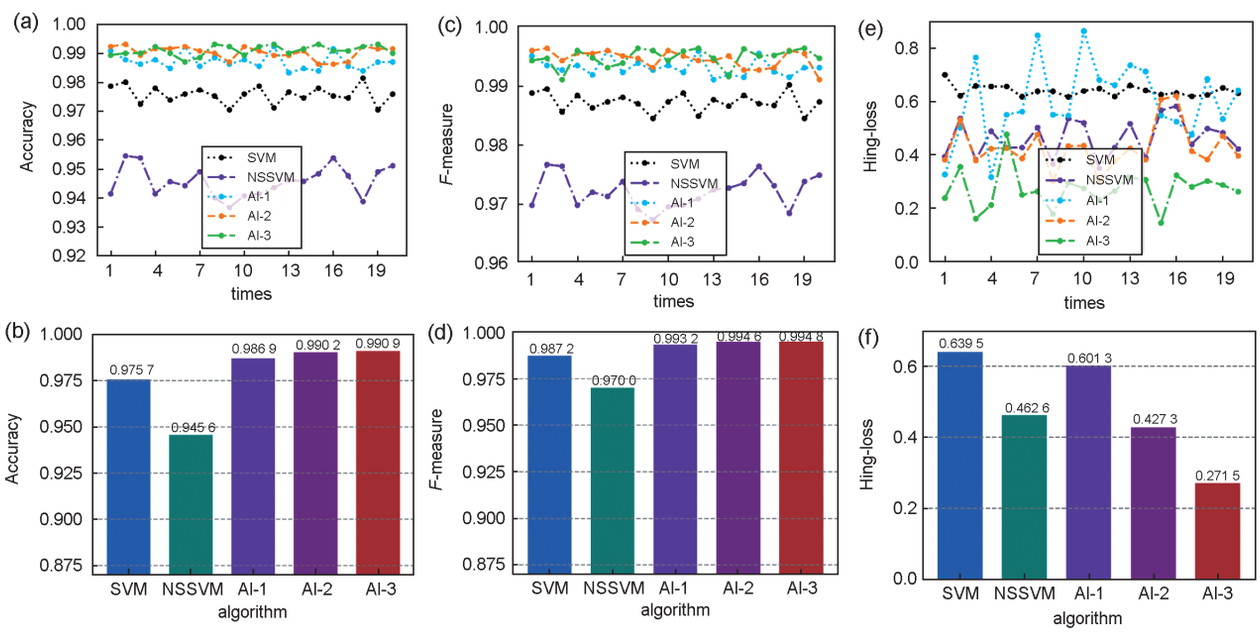
注:W表示方法1在数据集上的性能表法比方法2好;T表示两种算法性能相同;L表示方法1比方法2表现差。

4可以发现,与SVM和NSSVM相比,所提出的三种基于模糊隶属度的SVM样本约简算法确实产生了更好的分类性能。更重要的是,在约简了样本数量、减少了算法的训练时间和存储空间的同时,提高了SVM的分类性能。实验结果表明,最佳结果均来自于我们所提出的三种算法。就平均准确率和*F*-度量而言,Al-1、Al-2和Al-3算法分别产生了3个、1个和2个最佳

结果。就Hinge损失而言,AI-1、AI-2和AI-3算法分别产生了2个、1个和3个最佳结果。

为了进一步展现所提出算法的优越性,在表4中,将算法在不同数据集上的性能标准差进行标注。从表4可以看出,与SVM和NSSVM相比,我们提出的三种算法总体标准差更小,这意味着所提出的三种算法稳定性高,波动小。

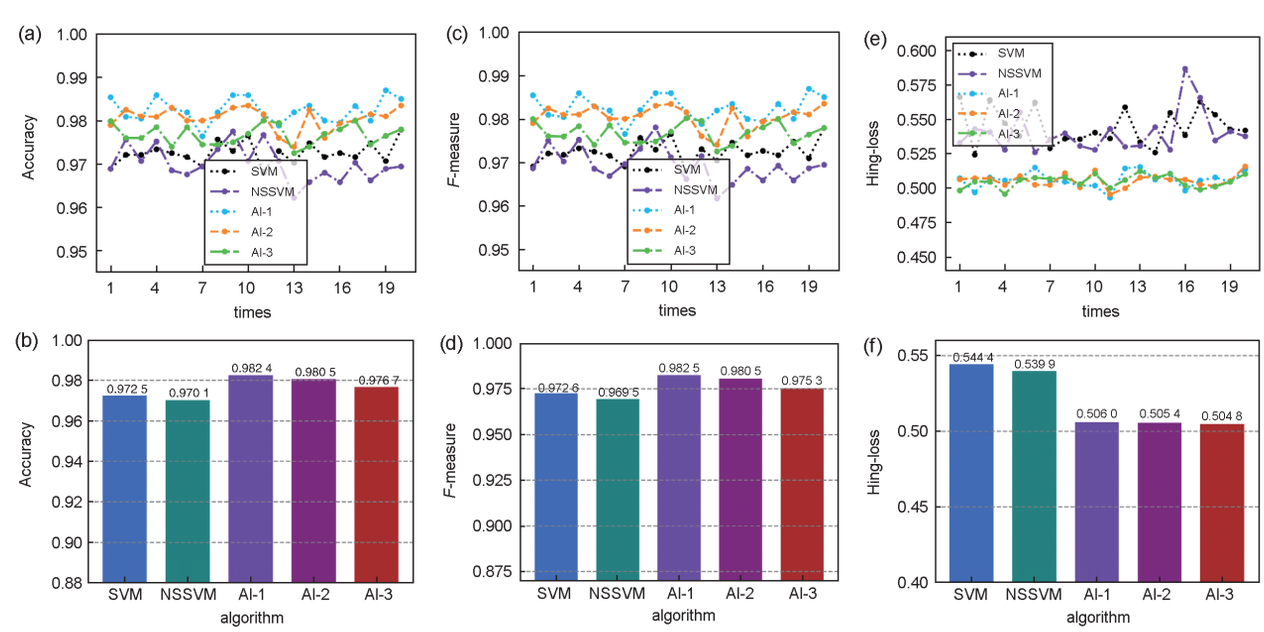
为了进行严格的对比,进行了配对t-检验[32]。配对t-检验用于分析比较同一个数据集上两个算法之间的差异是否显著。配对t-检验的p值表示两组比较结果来自平均值相等的分布的概率,p=0.05通常被认为具有统计学意义。表5列出了基于准确率、F-度量和Hinge损失的配对t-检验的结果。假设比较的两种方法(例如



注:(a) 准确率变化;(b) 平均准确率;(c) F-measure变化;(d) 平均F-measure变化;(e) Hinge损失变化;(f) 平均Hinge损失变化

图2 不同算法在wilt数据集上的性能对比

Fig. 2 Comparison of performance among different algorithms on the wilt data set



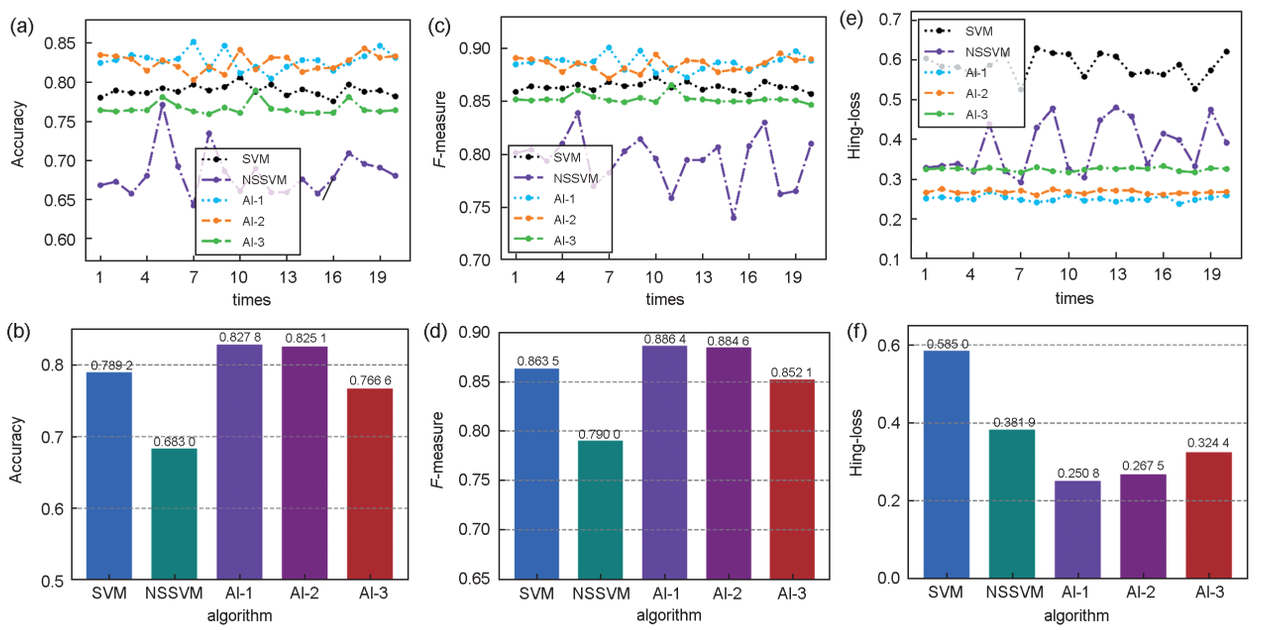
注:(a) 准确率变化;(b) 平均准确率;(c) F-measure变化;(d) 平均F-measure变化;(e) Hinge损失变化;(f) 平均Hinge损失变化

图3 不同算法在twonorm数据集上的性能对比

Fig. 3 Comparison of performance among different algorithms on the twonorm data set

方法1和方法2),W表示方法1在数据集上的性能表现比方法2好、T表示两种算法性能相同、L则表示方法1比方法2表现差。根据表5容易看出,就准确率而言,AI-1算法在6个数据集上的性能都优于其余对比算法;AI-2算法总体上优于对比算法,尤其在twonorm、titanic、mammographic和diabetes数据集上AI-2算法明显优于SVM算法;AI-3算法在titanic数据集上性能不如SVM,但在其余5个数据集上明显优于SVM。当比较F-度量和Hinge损失时,也可进行类似的分析。

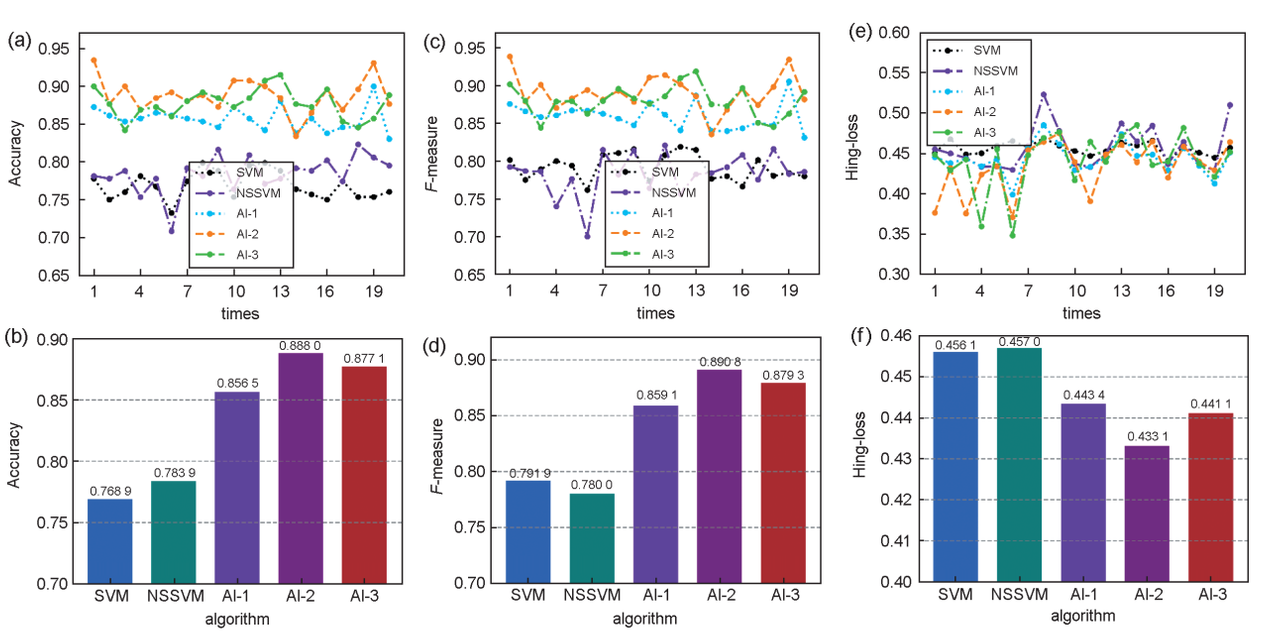
为更加清晰直观地展现所提出算法的优越性,图2至图7展示了20次重复实验后不同算法准确率、F-度量和Hinge损失的变化及其平



注:(a)准确率变化;(b)平均准确率;(c)F-measure变化;(d)平均F-measure变化;(e)Hinge损失变化;(f)平均Hinge损失变化

图4 不同算法在titanic数据集上的性能对比

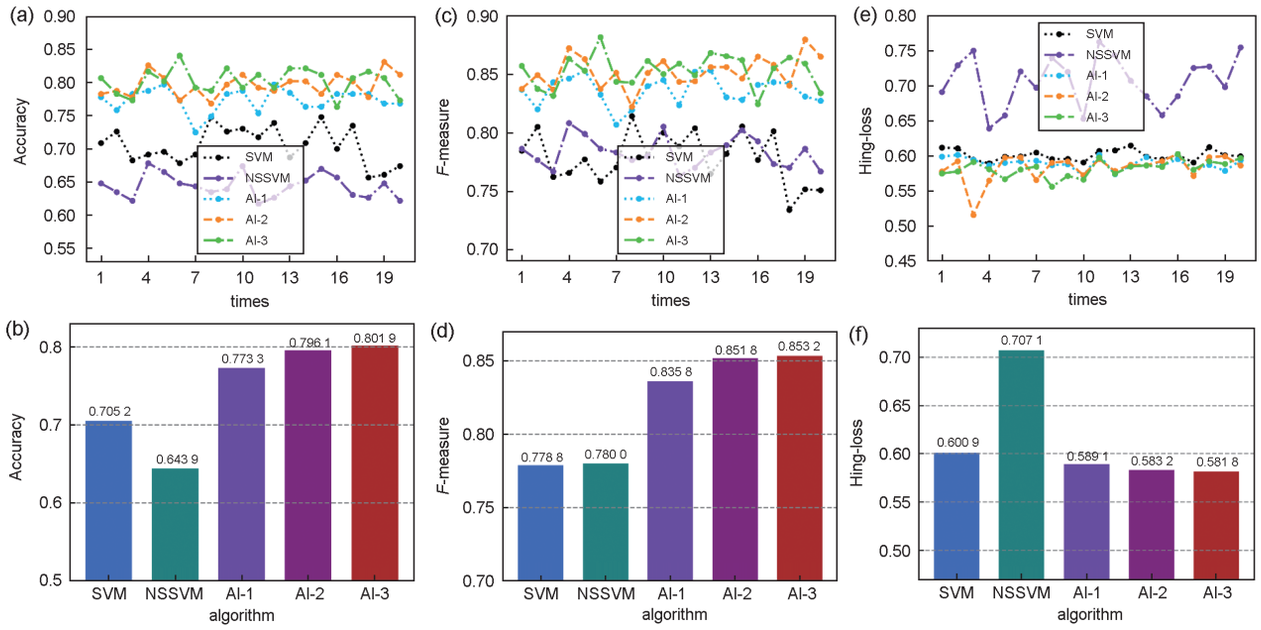
Fig. 4 Comparison of performance among different algorithms on the titanic data set



注:(a)准确率变化;(b)平均准确率;(c)F-measure变化;(d)平均F-measure变化;(e)Hinge损失变化;(f)平均Hinge损失变化

图5 不同算法在mammographic数据集上的性能对比

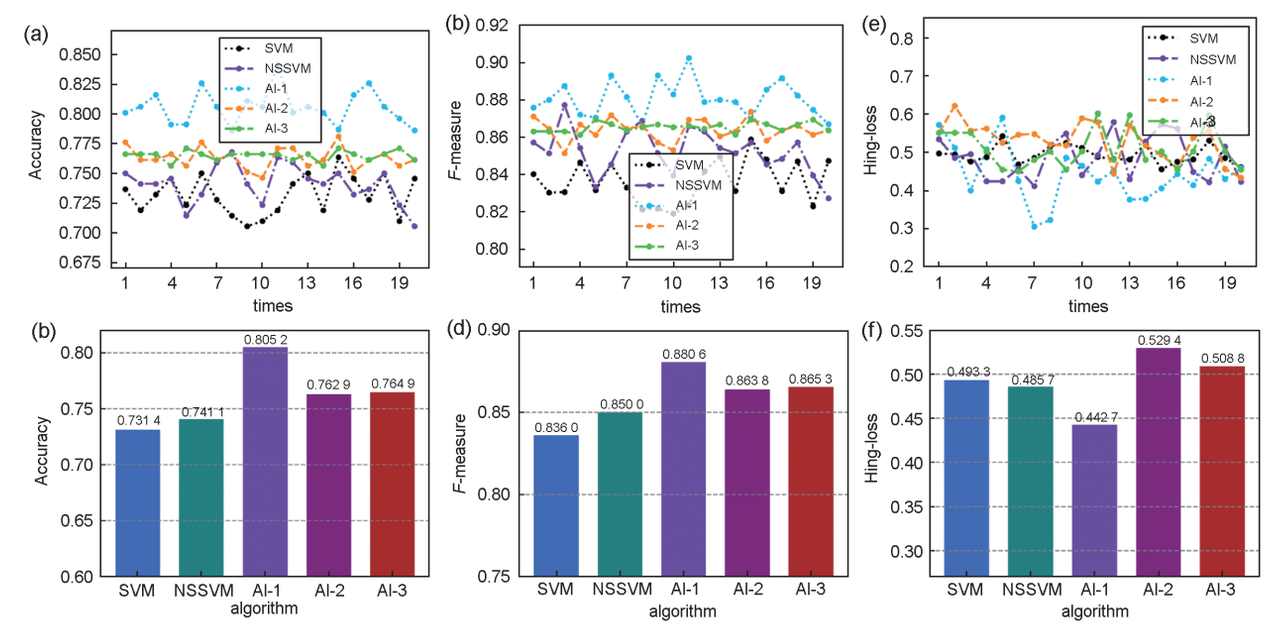
Fig. 5 Comparison of performance among different algorithms on the mammographic data set



注:(a) 准确率变化; (b) 平均准确率; (c) F-measure变化; (d) 平均F-measure变化; (e) Hinge损失变化; (f) 平均Hinge损失变化

图6 不同算法在diabetes数据集上的性能对比

Fig. 6 Comparison of performance among different algorithms on the diabetes data set



注:(a) 准确率变化; (b) 平均准确率; (c) F-measure变化; (d) 平均F-measure变化; (e) Hinge损失变化; (f) 平均Hinge损失变化

图7 不同算法在blood数据集上的性能对比

Fig. 7 Comparison of performance among different algorithms on the blood data set

均值。由图容易看出,我们所提出的算法在所选的6个数据集整体上实现了最高的准确率、F-度量和最小的Hinge损失,而且结果变化波动很小,实验结果很平稳。这意味着我们的算法在样本约简的同时,能够有效地去除噪声对实验的干扰,从而增强了训练过程的平稳性。

4 总结和展望

本文提出了一种基于模糊隶属度函数的SVM样本约简算法模型,根据不同的模糊隶属度函数,分别设计了基于类中心距离、KTA和CKA模糊隶属度函数的SVM样本约简算法。在所提出的模型中,其背后的算法思想是根据

不同的模糊隶属度函数计算样本的模糊隶属度值,再根据实验设定合适的模糊隶属度阈值,将模糊隶属度值低于阈值的样本进行约简,最后用于构建SVM样本分类器,以用于样本的最终预测。算法的关键是关于模糊隶属度函数的设计,不同分布情形的样本对应于不同的模糊隶属度函数,一个好的模糊隶属度函数将大大提高分类器的分类性能,对于大规模数据集,也将极大地减少样本的存储空间和计算时间。在多个数据集上的实验表明,基于模糊隶属度函数的样本约简算法是有效的。

未来的工作将集中于进一步验证所提出的方法在更真实的大规模数据中的应用,例如电商领域、医疗领域和交通领域。此外,将所提出的模型扩展到多核学习领域[33],并将其应用于多模态数据中[34]。最后,由于模糊隶属度阈值是根据样本的比例设定的,在真实的应用中可以根据样本的具体规模而设定,具有一定的灵活性,但不能保证该阈值是最佳的,在某种程度上也会影响SVM分类器的分类性能。因此,如何确定最佳的模糊隶属度阈值还需进一步研究。

参考文献:

[1] CORTES C, VAPNIK V. Support Vector Machine[J]. *Mach Learn*, 1995, **20**(3): 273−297. DOI: 10.1007/BF00994018.

[2] VAPNIK V N. An Overview of Statistical Learning Theory[J]. *IEEE Trans Neural Netw,* 1999, **10**(5): 988−999. DOI: 10.1109/72.788640.

[3] RAHMAN M A, HASAN S T, KADER M A. Computer Vision Based Industrial and Forest Fire Detection Using Support Vector Machine[C]//Proceedings of the 2022 International Conference on Innovations in Science, Engineering and Technology. 2022: 233−238. DOI: 10.1109/ICISET54810.2022.9775775.

[4] SINGH A, BANSAL A, CHAUHAN N, *et al*. Image Generation Using GAN and Its Classification Using SVM and CNN[C]//Proceedings of Emerging Trends and Technologies on Intelligent Systems. Springer Singapore, 2022: 89−100. DOI: 10.1007/978-981-16-3097-2_8.

[5] 潘华山, 严馨, 余正涛, 等. 基于支持向量机的越语新闻文本分类方法[J]. 山西大学学报(自然科学版), 2013, **36**(4): 505−509. DOI: 10.13451/j. cnki. shanxi. univ(nat. sci.). 2013.04.001.
PAN H S, YAN X, YU Z T, *et al*. Vietnamese News Text Classification Method Based on Support Vector Machine [J]. *J Shanxi Univ Nat Sci Ed*, 2013, **36**(4): 505−509. DOI: 10.13451/j.cnki.shanxi.univ(nat.sci.). 2013.04.001.

[6] LI Q, PENG H, LI J, *et al*. A Survey on Text Classification: From Traditional To Deep Learning[J]. *ACM Trans Intell Syst Technol*, 2022, **13**(2): 1−41. DOI: 10.1145/3495162.

[7] SU X, CAO C, ZENG X, *et al*. Application of DBN and GWO-SVM in Analog Circuit Fault Diagnosis[J]. *Sci Rep*, 2021, **11**(1): 1−14. DOI: 10.1038/s41598-021-86916-6.

[8] 梅御东, 陈旭, 孙毓忠, 等. 一种基于日志信息和CNN-text的软件系统异常检测方法[J]. 计算机学报, 2020, **43**(02): 366−380. DOI: 10.11897/SP.J.1016.2020.00366.
MEI Y D, CHEN X, SUN Y Z, *et al*. An Anomaly Detection Method for Software Systems Based on Log Information and CNN-Text[J]. *Chi J Comput*, 2020, **43**(02): 366−380. DOI: 10.11897/SP.J.1016.2020.00366.

[9] ALI W, TIAN W, DIN S U, *et al*. Classical and Modern Face Recognition Approaches: A Complete Review[J]. *Multi Tools Appl*, 2021, **80**(3): 4825−4880. DOI: 10.1007/s11042-020-09850-1.

[10] HE G, JIANG Y. Real-Time Face Recognition Using SVM, MLP and CNN[C]//Proceedings of the 2022 International Conference on Big Data, Information and Computer Network. 2022: 762−767. DOI: 10.1109/BDICN55575.2022.00149.

[11] DASH R, DASH D K, PANDA R S. Linguistic Information for Decision-Making Using SVM[M]//Advances in Data Science and Management. Springer, Singapore, 2022: 3−10. DOI: 10. 1007/978-981- 16-5685-9_1.

[12] ALSHUDUKHI J S. Smart and Interactive Healthcare System Based on Speech Recognition Using Soft Margin Formulation and Kernel Trick[J]. *Inter J Syst Assur Eng Manag* 2022: 1−10. DOI: 10.1007/s13198-022-01728-9.

[13] WANG Q. Support Vector Machine Algorithm in Machine Learning[C]//Proceedings of the International Conference on Artificial Intelligence and Computer Applications. 2022: 750−756. DOI: 10.1109/ICAICA54878.2022.9844516.

[14] DIVYANTH L G, CHELLADURAI V, LOGANATHAN M, *et al*. Identification of Green Gram (Vigna Radiata) Grains Infested by Callosobruchus Maculatus Through X-Ray Imaging and GAN-Based Image Augmentation[J]. *J Biosyst Eng*, 2022, **47**(3): 302−317. DOI:10.1007/s42853-022-00147-9.

[15] NAGPAL M, KAUSHAL M, SHARMA A. A Featurereduced Intrusion Detection System with Optimized SVM Using Big Bang Big Crunch Optimization[J]. *Wirel Pers Commun*, 2022, **122**(2): 1939−1965. DOI: 10.1007/s11277-021-08975-2.

[16] HASSANAT A B, ALI H N, TARAWNEH A S, *et al*. Magnetic Force Classifier: A Novel Method for Big Data Classification[J]. *IEEE Access*, 2022, **10**: 12592−12606. DOI: 10.1109/ACCESS.2022.3142888.

[17] KUDO T, MATSUMOTO Y. Chunking with Support Vector Machines[C]//Proceedings of the 2nd North American Chapter of the Association for Computational Linguistics. 2001: 1−8. DOI: 10.3115/1073336.1073361.

[18] SINGH K R, NEETHU K P, MADHUREKAA K, *et al*. Parallel SVM Model for Forest Fire Prediction[J]. *Soft Comput Lett*, 2021, **3**: 100014. DOI: 10.1016/j.socl.2021.100014.

[19] BADR E, ALMOTAIRI S, SALAM M A, *et al*. New Sequential and Parallel Support Vector Machine with Grey Wolf Optimizer for Breast Cancer Diagnosis[J]. *Alex Eng J*, 2022, **61**(3): 2520−2534. DOI: 10.1016/j.aej.2021.07.024.

[20] DESHPANDE N J, KARIBASAPPA K G, TOTAD S G. Comparative Analysis of Optimization Techniques on Multi-Class SVM[C]//Proceedings of the International Conference for Emerging Technology. 2022: 1−6. DOI: 10.1109/INCET54531.2022.9824946.

[21] ZHOU S. Sparse SVM for Sufficient Data Reduction [J]. *IEEE Trans Pattern Anal Mach Intell*, 2021, **44**(9): 5560−5571. DOI: 10.1109/TPAMI.2021.3075339.

[22] LEE Y J, MANGASARIAN O L. RSVM: Reduced Support Vector Machines[C]//Proceedings of the 2001 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2001: 1−17. DOI: 10.1137/1.9781611972719.13.

[23] GOODRICH B, ALBRECHT D, TISCHER P. Algorithms for the Computation of Reduced Convex Hulls [C]//Proceedings of the Australasian Joint Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2009: 230−239. DOI: 10.1007/978-3-642-10439-8_24.

[24] BANG S, JHUN M. Weighted Support Vector Machine Using K-Means Clustering[J]. *Commun Stat Simul Comput*, 2014, **43**(10): 2307−2324. DOI: 10.1080/03610918.2012.762388.

[25] CERVANTES J, LI X, YU W. Support Vector Machine Classification Based on Fuzzy Clustering for Large Data Sets[C]//Proceedings of Mexican International Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2006: 572−582. DOI: 10.1007/11925231_54.

[26] LIN C F, WANG S D. Fuzzy Support Vector Machines [J]. *IEEE Trans Neural Netw* 2002, **13**(2): 464−471. DOI: 10.1109/72.991432.

[27] JIANG X, YI Z, LV J C. Fuzzy SVM with a New Fuzzy Membership Function[J]. *Neural Comput Appl*,2006, **15**: 268−276. DOI: 10.1007/s00521-006-0028-z.

[28] LIN C, WANG S. Training Algorithms for Fuzzy Support Vector Machines with Noisy Data[J]. *Pattern Recognit Lett*, 2004, **25**(14): 1647−1656. DOI: 10.1016/j.patrec.2004.06.009.

[29] CRISTIANINI N, SHAWE-TAYLOR J, ELISSEEFF A, *et al*. On Kernel-Target Alignment[J]. *Adv Neural Inf Process Syst*, 2001, 14. DOI: 10.1007/3-540-33486-6_8.

[30] WANG T, QIU Y, HUA J. Centered Kernel Alignment Inspired Fuzzy Support Vector Machine[J]. *Fuzzy Sets Syst*, 2020, **394**: 110−123. DOI: 10.1016/j.fss.2019.09.017.

[31] HSU C W, CHANG C C, LIN C J. A Practical Guide to Support Vector Classification[J]. *Paediatr Perinat Epidemiol*, 2003. 1995, **9**(4): 455−467. DOI: 10.1111/j.1365-3016.1995.tb00168.x.

[32] DEMŠAR J. Statistical Comparisons of Classifiers Over Multiple Data Sets[J]. *J Mach Learn Res*, 2006, **7**: 1−30. DOI: 10.1007/s10846-005-9016-2.

[33] BAI Y X, LU X J. Multiple Kernel Learning-based Rule Reduction Method for Fuzzy Modeling[J]. *Fuzzy Sets Syst*, 2023, **465**(13): 108534. DOI: 10.1016/j.fss.2023.108534.

[34] JANGRA A, MUKHERJEE S, JATOWT A, *et al*. A Survey on Multi-Modal Summarization[J]. *ACM Comput Sur*, 2023, **55**(13s): 1−36. DOI: 10.1145/3584700.