计算机科学
COMPUTER SCIENCE

SVM样本约简算法研究综述

张代俐, 汪廷华, 朱兴淋

引用本文

张代俐, 汪廷华, 朱兴淋. SVM样本约简算法研究综述[J]. 计算机科学, 2024, 51(7): 59-70.

ZHANG Daili, WANG Tinghua, ZHU Xinglin. Overview of Sample Reduction Algorithms for Support Vector Machine [J]. Computer Science, 2024, 51(7): 59-70.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400143

SVM 样本约简算法研究综述

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
(17746670679@163.com)

摘　要　支持向量机(Support Vector Machine,SVM)是基于统计学习理论和结构风险最小化原则发展起来的一种有监督的机器学习算法,它有效克服了局部最小和维数灾难等问题,具有良好的泛化性能,并被广泛应用于模式识别和人工智能领域。但SVM的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,有时甚至无法执行的问题。为了缓解SVM在大规模训练集上存储需求高、训练时间长等问题,学者们提出了SVM样本约简算法。文中首先介绍了SVM理论基础,然后从基于聚类、几何分析、主动学习、增量学习和随机抽样5个方面系统综述了SVM样本约简算法的研究现状,讨论了各种SVM样本约简算法的优缺点,最后总结全文并展望未来。

关键词:支持向量机;大规模数据集;样本约简;机器学习;分类
中图分类号　TP181

Overview of Sample Reduction Algorithms for Support Vector Machine

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

Abstract　Support vector machine(SVM) is a supervised machine learning algorithm developed based on statistical learning theory and the principle of structural risk minimization,which effectively overcomes the problems of local minimum and curse of dimensionality and has good generalization performance. SVM has been widely used in the fields of pattern recognition and artificial intelligence. However,the learning efficiency of SVM decreases significantly with the increase of the number of training samples. For large-scale training datasets,the traditional SVM with standard optimization methods will be confronted with the problems of excessive memory requirements,slow training speed,and sometimes even being unable to execute. To alleviate the problems of high storage requirements and long training time of SVM on large-scale training sets,scholars have proposed SVM sample reduction algorithms. This paper firstly introduces the theoretical basis of the SVM and then systematically reviews the current research status of the SVM sample reduction algorithms from five aspects based on clustering,geometric analysis,active learning,incremental learning and random sampling,respectively. And it discusses the advantages and disadvantages of these algorithms,and finally presents an outlook on the future research of the SVM sample reduction methods.

Keywords　Support vector machine,Large-scale data set,Sample reduction,Machine learning,Classification

1　引言

随着科技的发展,人们需要处理的信息数据呈现出高维和海量的特点,然而,如何有效地分析和利用这些数据,成为了模式识别、数据挖掘、神经网络、机器学习等学科共同面临的问题[1]。在统计模式识别中,许多分类方法的计算复杂度随着训练集样本个数的增加而快速增长,因此对于较大规模数据的有效分类和解释变得更具挑战性[2]。各种机器学习算法,例如K近邻(K-Nearest Neighbor,KNN)[3-5]、决策树(Decision Tree)[6-7]和支持向量机(Support Vector Machine,SVM)[8-9]常用于现实场景的数据分类。一般来说,解决目标问题的数据越多,对结果有用的信息就越准确可靠。然而,在设备资源受限的物联网环境中,海量数据的存在大大增加了服务器数据管理的存储与计算开销,同时也容易导致通信网络陷入拥堵状态[10]。因此,在不显著降低数据处理效果的情况下减少数据量至关重要[11]。

机器学习算法可分为3类[12-13]:有监督学习、无监督学习和半监督学习。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维理论和结构风险最小原理,很大程度上克服了

到稿日期:2023-04-24　返修日期:2023-09-06
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)
This work was supported by the National Natural Science Foundation of China(61966002) and Graduate Innovation Found of Jiangxi Province (YC2022-s944).
通信作者:汪廷华 (wthpku@163.com)

局部极小和维数灾难问题[14]。由于SVM具有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[15-16]、文本分类[17-18]、故障诊断[19-21]、面部识别[22-23]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[24]，而分类超平面仅由支持向量(Support Vector，SV)决定[25-26]。SVM适用于线性可分和非线性可分的数据分类。从几何的角度来看，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[27]。针对现实任务中的非线性可分的数据集，SVM引入核技巧[28-30]，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，训练时间和计算开销显著增加。

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的增加呈平方增长[31]。另外，凸二次规划的优化计算时间也会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 $O(t^3)$[32]，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，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，有时甚至无法执行的问题[33-34]。SVM的分类原理给予了学者们这样的启迪：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。样本约简可以根据样本的重要程度选择训练样本集，尽可能保留对分类超平面有决策作用的训练样本作为候选支持向量集，并剔除非支持向量以缩小训练集，最后将候选支持向量集用于SVM分类器进行训练。样本约简既可以起到降低算法计算代价、加快学习速度的作用，也可能避免“过拟合”现象[35]的发生，从而提高分类算法的泛化能力。

如今，由于移动互联网和云计算的快速发展，全球数据呈现出爆炸式增长的趋势。而大型数据的存储和计算将导致传统的SVM分类器运算速度和学习效率显著降低。为使SVM在当今时代更好地发挥其良好泛化性能的优势，对于大规模数据的约简是极其必要的。针对不同的样本选择策略，国内外学者们提出了不同类型的样本约简算法，这对于在大数据时代使用SVM分类器具有重要意义，但国内尚未有针对此研究领域的相关综述。鉴于此，本文综述了SVM样本约简方法的研究现状，详细论述了现有的各种SVM样本约简方法，并阐述了各种方法背后的思想及其优缺点，给SVM样本约简算法的研究与发展提供了丰富的设计思路和广泛的应用前景。在分析各种样本约简算法的基础上，进一步凝练其研究方向。

2 支持向量机理论基础

假设训练样本集 $\boldsymbol{T}=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_t,y_t)\}$ 由 t 个训练样本 $\boldsymbol{x}_i$ 组成，其中 $\boldsymbol{x}_i\in R^d$，$i\in 1,\cdots,t$，以二分类问题为例，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。最优超平面可通过求解以下优化问题得到：

$$\begin{aligned}&\max_{\boldsymbol{w},b,\xi_i}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{t}\xi_i\\&\text{s.t. }\ y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\geqslant 1-\xi_i\\&\qquad\xi_i\geqslant 0,i=1,2,\cdots,t\end{aligned}\tag{1}$$

其中，ξ_i 为松弛变量，C 是权衡控制边界和松弛惩罚之间的参数。为解决上述SVM优化问题，本文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：

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

为避免在高维特征空间中计算困难，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 ϕ，而不必显式计算出 ϕ 的具体值。

SVM的决策函数可以被定义为：

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

由式(1)和式(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，其将变得不可行。尽管存在旨在加速SVM训练的技术，其中包括chunking[36]、并行[37-38]和修改工作集[39]的方法，但它们通常在优化过程中引入了额外的内存负担。因此，减少SVM训练集大小的算法被认为是解决从大型数据集中学习SVM问题的直接有效措施。

SVM分类时间依赖于支持向量的数量，只有 $\alpha_i>0$ 的向量才有助于决策。因此，支持向量的数量应保持在较低水平。此外，支持向量的数量间接取决于训练集的大小 t。$\boldsymbol{T}$ 中的样本数量越少，在训练过程中确定的支持向量就越少。因此，在训练开始之前，对训练集进行预处理，筛选出对分类超平面有决策作用的支持向量，剔除非支持向量，这对减少样本训练时间和存储空间至关重要。

3 SVM样本约简算法

本章系统分析了SVM样本约简算法，根据不同的样本约简策略，产生了不同的约简算法。本文将这些样本约简算法主要分为5个类别，具体分类如图1所示。

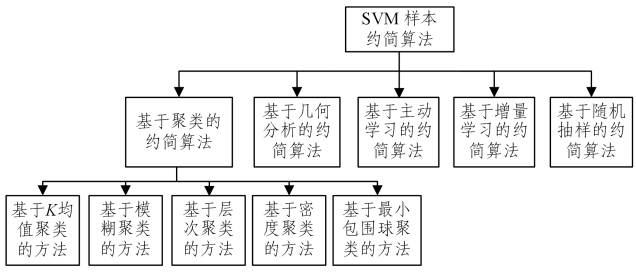

图1 SVM样本约简算法分类

Fig. 1 Classification of SVM sample reduction algorithms

3.1 基于聚类的约简算法

传统的聚类是一种无监督分析方法,在对数据进行划分的过程中不依赖任何先验知识和背景假设,单纯地按照相似性原则进行自然划分,让每一类的数据关系尽可能相似,而类与类之间的关系不相似[40]。广泛的聚类算法包括 K 均值聚类、模糊聚类、层次聚类、密度聚类和最小包围球聚类,已被用于 SVM 分类以减小训练集的大小。

3.1.1 基于 K 均值聚类的方法

K 均值聚类算法是基于数据之间相似性对数据进行聚类的强大聚类算法之一。算法基本思想是将数据集按照不同的类别划分成多个簇,使得数据点和相应的簇中心的欧氏距离最小,其具体步骤如下:(1)从 t 个数据中随机选择 p 个数据对象作为初始聚类中心$\{m_i\}$;(2)分别计算每个数据点到 p 个初始聚类中心的欧氏距离,根据距离最小原则重新分类该数据;(3)计算每个聚类簇的平均值,得到新的聚类中心。重复步骤(2)和(3),直到聚类中心不再改变或者聚类中心发生偏移的距离小于设定的阈值[41]。该簇的聚类中心可由式(4)计算得出:

$$m_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_j \tag{4}$$

其中 n_i 为簇内样本个数。

$$\|\boldsymbol{x}-m_i\|<\|\boldsymbol{x}-m_j\| \tag{5}$$

在每次迭代时,根据式(5)更新数据 $\boldsymbol{x}$ 的所属簇,若数据 $\boldsymbol{x}$ 满足式(5),则将其分配给簇 m_i,反之,则将其分配给簇 m_j。

Yao 等[42]提出的方法旨在使用聚类算法选择训练样本。该算法通过在原始训练集上运行聚类算法,使每个样本都获得一个聚类标签,然后将其与真实标签进行比较,若两者不一致,则该样本可能位于两个簇的边缘区域,并将其视为误分类样本。然后,在这些误分类的样本中,选择其附近最近的几个邻居,将它们用于训练。该算法大大减小了训练集的规模,从而有效地节省了 SVM 的训练和预测时间,同时保证了分类器的泛化性能。Almeida 等[43]将 K 均值聚类应用于 SVM 样本约简,选择一小部分训练数据点用于训练 SVM。这种被称为 SVM-KM 的方案中的数据点被聚类成几个簇,簇内样本包括单类标签和多类标签。包含单个类标签的数据点的簇仅保留簇中心,而簇中的其他数据点被删除,具有多个类别标签的簇保持不变,保留簇内所有训练样本。该算法不仅减少了训练样本的数量,加快了 SVM 的训练速度,而且在不影响 SVM 泛化能力的情况下降低了计算复杂度。但是 SVM-KM 算法的性能受数据的分布和簇数目的影响,并且对训练数据集的数量非常敏感。此外,SVM-KM 在处理低维密集训练集时是有效的,可以减小训练集的大小,但当维度变大且数据稀疏时,分类的准确性会受到影响。为了提高 SVM-KM 的准确性,考虑到不同聚类中心对分类的贡献程度不同,Bang 等[44]提出了 WKM-SVM 改进算法,该算法通过将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,该算法不仅解决了 KM-SVM 分类精度低的问题,且能有效应用于处理不平衡数据集。

Koggalage 等[45]也应用 K 均值算法,通过构建可变半径识别"纯簇"(Crisp Cluster)。"纯簇"是具有相同标签样本的簇,"非纯簇"是包含异类标签的簇。该算法考虑到若"纯簇"在靠近边界处,则"纯簇"也可能在簇的边界中包含支持向量。该算法通过簇内半径和簇内样本数量构建簇内安全区域,形成内部簇和外部簇,其中内部簇的样本远离决策边界。该内部簇中的样本是要移除的检测样本,内部簇和外部簇之间的样本不得移除。针对可变半径的确定,该文献提出了基于初始簇和基于聚类中心两种算法。该算法最终保留细化后的"纯簇"和"非纯簇"用于 SVM 训练,减小了训练样本的规模。

Shen 等[46]在 K 均值聚类算法的基础上引入了 FIFDR(Fast Iteration of Fisher Discrimnant Ratio,FIFDR)算法,该算法首先根据数据点的类别标签将聚类进一步划分为更小的聚类,然后通过将簇的质心作为训练数据来获得近似超平面,并且使用最大-最小聚类距离算法(Max-Min Cluster Distance,MMCD)去除远离近似超平面的冗余簇;其次使用 FIFDR 算法,以去除每个剩余簇中不必要的数据点。该算法假设位于聚类重心周围的聚类数据点是密集数据点,位于簇的内层,这些数据点被视为没有包含支持向量并被删除。而分散的数据点是簇外部层中的稀疏数据点,这些数据点被视为包含支持向量,因此被保留。Fisher 判别比(Fisher Discrimnant Ratio,FDR)用于确定一个簇中的密集和分散数据点之间的边界,该边界是根据数据点到聚类质心的距离密度计算的。该算法在大规模数据集的分类任务中,训练时间显著减少,而精度仍然很高。

3.1.2 基于模糊聚类的方法

模糊聚类是另一种聚类算法,与 K 均值算法不同的是,模糊聚类允许将某个数据点分配给多个簇[47]。对于某一个样本,不再是将其赋予某一类标签,而是通过一个隶属度函数计算隶属度值,根据隶属度值进行分配。隶属度值表示该样本与每个簇的相似度,相似度越高,则其属于该簇的概率越大。Cervantes 等[48]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比传统的 SVM 更短,而且获得的支持向量数量相似。

Manimala 等[49]使用模糊 C 均值(Fuzzy C-Means,FCM)算法对训练数据进行聚类,保留由每个簇的质心组成的数据集,并根据簇内数据与质心之间的距离选择边界样本。该算法与传统的 SVM 训练算法相比,减少了分类时间并提高了准确性,但是该算法中的聚类数和每个类的选择样本数是重要参数,参数的选择会显著影响算法的性能。文献[50]应用 CH 算法和模糊聚类对数据点进行聚类,仅选择每个聚类的中心作为训练数据点。该算法可以识别包含噪声或可能有噪声的数据,提高了 SVM 分类器的泛化性能。文献[51]提出了一种新的两步样本约简算法,在算法的实现中,首先使用 FCM 算法对原始训练样本进行聚类,然后使用多元高斯分布(Multivariate Gaussian Distribution,MGD)模型通过选择下一次训练的边界样本来减少训练样本,MGD 模型可以将样本转换为概率分布函数的值。通过观察发现,靠近高斯分布中心的样本具有较大的概率分布函数值,而靠近边界的样本具有

较小的概率值，因此可以通过其函数值来选择数据。实验表明，该算法在不降低分类精度的情况下加快了训练速度。

3.1.3 基于层次聚类的方法

层次聚类是另一种降低支持向量机在大规模数据集中复杂性的聚类算法。Krishna 等[52]提出了一种使用 BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) 算法的层次聚类方法，通过采用 BIRCH 层次聚类进行数据预处理。BIRCH 层次聚类可以为 SVM 训练提供高质量、抽象和简化的数据集，而不是原始的大型数据集。因此，除了显着减少训练时间外，生成的 SVM 分类器比使用原始冗余数据集的 SVM 分类器表现出了更好的性能。文献[53]提出了一种新的可扩展和可靠的 SVM 分类方法，称为基于层次聚类的支持向量机(Clustering-Based SVM，CB-SVM)，其为 SVM 分类器提供具有价值的样本。CB-SVM 通过将输入的训练集 $\boldsymbol{T}$ 添加到簇中来构建聚类特征 CF 树(CF Tree)。该算法采用分层微聚类算法[54]，只扫描原始训练集一次。它不允许回溯，因此数据分布可能会影响其能力，但 CF 树仍然可以提取主要数据分布模式。聚类特征(对于任意的簇 c_i)通过式(6)的三元组给出：

$$CF=(t_i,LS,SS) \tag{6}$$

$$LS=\sum_{j=i}^{t_i}\boldsymbol{x}_j \tag{7}$$

$$SS=\sum_{j=i}^{t_i}\boldsymbol{x}_j{}^2 \tag{8}$$

其中，t_i 为簇内样本个数。

CF 树是高度平衡的树，其特征在于两个参数：分支因子(b_{CF}) 和阈值(t_{CF})。分支因子定义了每个非叶节点孩子的最大数目，而阈值给出了存储在叶子节点中子聚类的最大半径，这两个参数影响着 CF 树的大小[55-56]。CF 树是按照类似于 B^+ 一树中应用的程序构建的。这种聚类有一个值得注意的特征和优势是其可以处理异常值和噪声数据，那些包含明显少于其他向量数量的叶子被视为异常值。同时，CB-SVM 的训练复杂度二次依赖于支持向量的数量，通常比整个数据集的训练复杂度小得多。此外，CB-SVM 对于规模非常大的数据集具有高度的可扩展性，并且分类非常准确。但是在高维特征空间中它目前仅限于使用线性内核，还需谨慎选择算法中的 b_{CF} 和 t_{CF} 两个参数。Awad 等[57]提出的算法不使用原始数据来训练 SVM，而是使用层次聚类算法生成的树节点，使用动态增长自组织树(Dynamically Growing Self-Organizing Tree，DGSOT) 算法[58]产生的层次结构来近似支持向量。由于层次树的大小明显小于原始数据集的大小，因此训练过程将加快。

3.1.4 基于密度聚类的方法

密度聚类方法是经典聚类算法的一个重要分类，它能够发现具有相同密度结构的数据，而不拘泥于数据的凹凸类型和聚类形状[59]。Wu 等[60]根据密度聚类能找到代表聚类形状的边缘点集的特点，提出了基于密度聚类的支持向量机分类算法(SVM-DC 算法)，同时，通过控制核心点邻域内噪声点的比例，添加可能成为支持向量的样本，使之能够更精确地完成支持向量机的分类任务。Zhang 等[61]提出了一种改进的基于密度聚类的模糊支持向量机算法(DBSCAN)，运用 DBSCAN 算法对原始训练集进行处理，去除对分类超平面贡献小的中心样本，用剩余的边缘样本集完成训练。实验表明，该方法形成的聚类边缘较好地保持了原样本的分布情况，在保证分类精度的同时，缩短了训练时间。但是算法中需要人为给定同类样本阈值 θ、核心点纯度 η 和 ε-邻域几个参数，由于参数的选取往往依赖于数据分布的具体特性，当用户缺乏有关数据分布的先验知识或者数据维数较大时，难以设置合理的参数。针对文献[61]中参数设置的问题，Akasapu 等[62]提供了一种基于相对密度的聚类算法，有效解决了 DBSCAN 对用户自定义参数非常敏感的问题。文献[63]运用 OPTICS 算法能发现任意形状的聚类，且具有对输入参数不敏感的优势，提出了一种基于 OPTICS 密度聚类的支持向量机算法，通过对原始数据进行预处理，利用可达图得到约简样本代替原始训练样本进行训练，降低了 SVM 分类器训练所需的时间和空间复杂度。

3.1.5 基于最小包围球聚类的方法

最小包围球聚类(Minimum Enclosing Ball，MEB) 最初用于对 SVM 中的 VC 维进行估计[64-65]。这种分类算法具有计算速度快、鲁棒性强等优点[66]。Cervantes 等[67]引入了 MEB 的概念，MEB 被定义为通过核心集思想围绕给定集合的最小球，其基本思想是寻找一个超球 $B(c_B,r_B)$(c_B 和 r_B 分别为超球的中心和半径)，半径 r_B 需尽可能小且包围尽可能多的数据点。由于为给定集合找到最佳包围球非常具有挑战性，因此 Cervantes 等建议使用$(1+\sigma)$-近似的 MEB，即对于给定的 $\sigma>0$，球$(c_B,(1+\sigma)r_B)$是 MEB 的一个$(1+\sigma)$近似。在 MEB 聚类之后，一个细化的训练集包含来自混合类标签聚类的所有向量，以及仅包含同类标签的聚类的质心。在 SVM 训练之后，应用额外的 de-clustering 技术来恢复位于决策超平面附近的其他潜在有价值的样本，并将它们附加到候选训练集中。与其他方案相比，MEB 中不需要最优聚类数。不是支持向量的球由 SMO 算法确定并从训练数据集中删除，剩余球的数据点用于找到分类超平面所需的最终支持向量。该方法以略微降低分类准确度为代价减少了训练时间。

针对核心集的选取，Tsang 等[68]利用“近似性”进行了拓展，提出了核心支持向量机(Core Vector Machine，CVM) 算法，通过计算几何中的 MEB 问题，找到违反 KKT 条件且与核心集中心的距离最大的样本，将其添加到核心集，同时更新核心集的中心和半径，直到训练集中没有违背 KKT 条件的样本。在大型数据和现实世界数据集上的实验表明，CVM 与现有 SVM 一样准确，并且比传统的的 SVM 方法快得多，可以处理更大的数据集，其时间复杂度与样本数量呈线性关系，而空间复杂度与样本数量无关。但由于核心集上的向量是增量添加的，且从未移除，因此可能存在冗余的样本，其次，CVM 只能用于特定的某些核函数和核方法。

文献[69]引入了中心约束 MEB 问题，随后扩展了 CVM 算法，提出了广义 CVM 算法(Generalized Core Vector Machines，GVM)。该算法继承了 CVM 算法的简单性，其时间复杂度也是线性的，且空间复杂度与时间无关。并且 GVM 算法可以用于任何线性或者非线性核上，在大型数据集上比 CVM 算法表现得更灵活，在生成更小规模约简样本的同时

没有降低分类准确率。针对 CVM 算法存在样本冗余的缺陷,文献[70]提出了一种基于有效冗余样本删除技术的在线 CVM 分类算法,该算法能够自适应调整 MEB,可以实现分类器的在线更新,安全且永久地删除大多数冗余样本。

3.1.6 基于聚类的约简算法小结

各种基于聚类的 SVM 样本约简方法在实际应用中的效果不同。例如,基于 K 均值聚类的 SVM 样本约简算法简单易收敛,适用于大规模凸结构的训练集。Bang 等[44]在 diabetes 数据集上的实验结果表明,与传统的 SVM 分类算法相比,WKM-SVM 算法的分类错误率降低了 1.7%,分类时间减少了 55%。这也从另一方面说明,基于 K 均值聚类的 SVM 样本约简算法极大地提高了算法的分类效率。但是该算法易受噪声和初始样本的影响。针对其易受噪声影响的问题,Manimala 和 Yu 分别提出了基于模糊聚类和层次聚类的 SVM 样本约简算法,这类算法具有很强的抗噪能力。Cervantes 等[48]在 waveform 数据集上的实验结果表明,所提出的基于模糊聚类的 SVM 约简算法在约简了 25%的原始数据时,准确率比传统的 SVM 算法提高了 0.2%,但是初始样本和初始参数容易影响该算法的性能。为此,学者们提出了基于密度聚类和最小包围球聚类的 SVM 样本约简算法,该类算法能够自适应聚类,不需要选择最佳聚类数,例如,Wu 等[60]在 adult 数据集上的实验结果表明,采用基于密度聚类的 SVM-DC 约简算法可以使 SVM 的分类准确率从 86.81%提高到 95.43%。除此之外,核函数的计算数量级也由原来的 10^9 降至 10^6 以下。这也说明,基于密度聚类的约简算法能够显著降低算法的时间复杂度。总的来说,还需要针对具体的样本分布情况来确定合适的聚类约简算法。

3.2 基于几何分析的约简算法

SVM 的本质是在两类数据之间寻找最优超平面,从训练样本的几何分布上看,最有可能成为支持向量的样本分布在边界上或者靠近边界的区域。凸包能刻画数据样本的局部分布,数据集 $\boldsymbol{T}$ 的凸包 $CH(\boldsymbol{T})$ 通过式(9)定义:

$$CH(\boldsymbol{T})=\{\sum_{i=1}^{t}\tau_i\boldsymbol{x}_i \mid \boldsymbol{x}_i\in\boldsymbol{T},\sum_{i=1}^{t}\tau_i=1,\tau_i\geqslant 0\} \tag{9}$$

壳向量指所有位于训练集最边缘的样本,即位于训练集凸包上的样本。对于线性可分数据集,支持向量集就是壳向量的子集[71],壳向量往往只占数据集的很小一部分。因此,使用壳向量作为新的训练集可以提高 SVM 的训练速度。

对于线性可分训练集,文献[72]构造包含每个类的最小凸包,并在凸包中找到最近点,连接最近点之间的线段,选择垂直平分该线段的超平面作为最优分类超平面。文献[73-77]针对线性不可分的情形,提出了约简凸包(Reduced Convex Hull,RCH)算法。RCH 与凸包的区别在于引入了上界因子 δ,约简凸包通过式(10)定义:

$$RCH(\boldsymbol{T},\delta)=\{\sum_{i=1}^{t}\tau_i\boldsymbol{x}_i \mid \boldsymbol{x}_i\in\boldsymbol{T},\sum_{i=1}^{t}\tau_i=1,0\leqslant\tau_i<\delta\} \tag{10}$$

其中,$0<\delta\leqslant 1$。由此可知,$RCH(\boldsymbol{T},\delta)\subseteq CH(\boldsymbol{T})$。$\delta$ 减小,约简凸包非均匀地向质心收缩,使得更多的数据点离开凸包。当 $\delta=\frac{1}{t}$ 时,约简凸包仅包含质心,当 $\delta=1$,$RCH(\boldsymbol{T},\delta)=CH(\boldsymbol{T})$。因此可以通过改变 δ 的大小来减少异类凸包的重叠,将线性不可分转化为线性可分。Mavroforakis 等[77]提出了 RCH-SK 算法,该算法通过计算 RCH 中的极值点在特定方向上的投影值,选取投影值最小的数据用于训练。该算法可以直接用于线性不可分的数据集。受 RCH 算法的启发,由于 RCH 的复杂性随着约简因子 δ 变小而增加,且极值点的数量和 RCH 的形状随约简因子 δ 的变化而变化,因此,文献[78]提出了一种 SCH-SK 算法。该算法更容易计算极值点在特定方向上的投影,同时候选极值点考虑了每类数据分布,因此该算法具有良好的泛化能力。尽管通过在 RCH 方法中为每个类选择合适的缩减因子获得了可接受的泛化性能,但没有确定的方法可以选择最优缩减因子。

Chau 等[79]引入了凹凸包(Convex Concave Hull,CCH)的概念,提出了一种 CCHSVM(Convex Concave Hull SVM)算法。该算法首先找到数据集的凸包,然后使用 K 近邻算法寻找凸包极值点的 K 个最近数据点,用于构造 CCH。最终的训练数据集由原始的凸包顶点和靠近凸包顶点边界的凹凸包数据点构成。实验结果表明,CCHSVM 方法获得的精度与经典 SVM 方法非常接近,训练时间显著缩短,但是 k 的值会影响 CCH 的外观[80]。文献[81]提出的基于聚类的凸包算法(Clustering Based Convex Hull,CBCH)进一步降低了 SVM 处理大规模训练集的复杂性。CBCH 算法首先运用 K 均值聚类算法对训练集进行聚类,再使用 Quickhull[82]算法获得包含同类标签的簇的凸包,只有凸包的顶点和包含多个类别标签的簇最终被保留,用于 SVM 训练。该算法适用于线性可分和线性不可分数据集,可以有效地去除冗余数据而不降低精度,从而减少了分类时间。但由于 K 的值是影响 CBCH 算法的重要因素,若 K 值过大,则最后保留的数据点会过多,导致计算机的开销增加。Wang 等[83]提出了一种在线 SVM 分类算法(VS-OSVM 算法),该算法是基于每类中凸包顶点选择数据构成近似凸包,在线分类过程中,SVM 根据选择的样本和新到达的样本进行在线更新。该算法不能消除新到达的噪声数据样本,其次,初始数据集的选取对该算法的分类有影响。由于初始数据集不能代表整个数据集的分布,最终真实分类器的一些重要支持向量可能会在在线更新过程中被删除,因此用所选样本训练的初始分类器可能会导致与最终真实分类器的偏差较大。

对于非线性可分数据,文献[84]引入核函数,使得该算法可以直接用于非线性情况,并且该算法对内存存储的要求在数据方面是线性的,也适用于大型训练集。文献[85]引入了一种新的随机近似凸包算法,该算法可以在可接受的执行时间内用于高维,且内存需求较低。文献[82]是以随机增量算法为基础进行改进的算法,能够在高维空间上应用。该算法通过大幅度排除非凸包顶点,最终形成凸包用于训练,算法的平均复杂度为 $O(t\log t)$。Osuna 等[86]提出了一种在特征空间中寻找数据点的凸包极值点的算法,该方法的关键是使用核函数和大型数据集的增量过程在特征空间中构造凸包。该算法的一个重要缺点是凸包的复杂性和极值点的数量与特征空间的维数呈指数关系,因此该方法在高维数据上的适应性有待进一步研究。

基于几何分析的约简算法充分结合数据的几何分布,

考虑了边界区域上的样本更有可能是支持向量的特性，从而能够快速去除距离分类边界较远的样本，具有计算量小、内存需求低等优点。与传统的 SVM 算法相比，Chau 等[79]在 checker-board 数据集上的实验结果表明，在获得同等准确率的情况下，基于几何的 CCHSVM 约简算法支持向量的个数减少了一半，训练时间减少了 72%。但基于几何分析的约简算法容易受噪声和离群点的影响，处理不平衡数据的效果不是很理想，算法性能还有待提升。

3.3 基于增量学习的约简算法

增量学习是对新增训练集进行不断更新学习的算法，其思想充分考虑了新增训练集对原有训练集的影响，同时充分利用了历史分类的结果，使得学习过程具有延续性。

支持向量机的增量学习研究始于 Syed 等[87]的研究工作，该算法将数据分批次进行学习，在每次学习中只保留支持向量，丢弃所有的非支持向量；对于每一批新数据，将之前保留下的支持向量作为候选支持向量集用于 SVM 训练，依次选择新一批数据中的支持向量。该算法极大地减少了原始训练集的样本数量，提高了训练速度，但分批次训练对计算机的内存需求高，并且，如果要用候选集支持向量集代替原始训练集进行训练，则需要考虑数据集的分布问题，如果参与训练的候选支持向量集与新一批参与训练的样本的差异很大，那么该算法的准确率将会很低。考虑到采样方法对增量 SVM 学习能力的影响，Xu 等[88]提出了基于马尔可夫重采样的的增量 SVM 算法(Incremental Support Vector Machine Based on Markov Resampling，MR-ISVM)，结果表明，该算法缩短了采样和训练的总时间，并且误分类准确率更低。文献[89]提出了一种基于压缩的 k 近邻增量学习算法，该算法采用主动增量学习和 k 近邻算法提取位于类边界附近的数据点，避免了以批处理方式训练 SVM 需要大量内存的问题。文献[90]提出了增量学习的精确解，即增加一个训练样本或减少一个训练样本对拉格朗日系数 α_i 和支持向量的影响。除此之外，该算法利用增量训练前距离分类超平面的距离来设置删除数据的区域。但如果存在超平面随着增量训练旋转的问题，那么远离超平面的数据又可能成为支持向量，因此该方法对于分离超平面的旋转适应性相对比较脆弱。为了使增量学习算法对分类超平面的旋转具有鲁棒性，Katagiri 等[91]提出了基于一类支持向量机的增量学习算法，该算法为每个类生成一个超球面，保留存在于超球面边界附近的数据作为支持向量的候选，并删除其他数据。实验结果表明，该算法可以在不降低 SVM 泛化能力的前提下删除大量数据。

文献[92]从理论和实验两个方面证明了当前增量学习 SVM 算法存在丢失增量样本信息的问题，提出了基于超平面距离的支持向量机增量学习算法(Hyperplane-Distance-SVM，HD-SVM)。该算法根据支持向量的几何特性，利用超平面距离提取样本，选取最有可能成为支持向量的样本组成边界支持向量集，并对边界支持向量集进行 SVM 训练。该方法减少了训练样本的数量，有效地提高了增量学习的训练速度。文献[71]考虑到新增样本可能带有原样本集不包含的分类信息，将增量学习与几何分析相结合，提出了基于壳向量的线性支持向量机的快速增量学习算法。该算法首先求出初始样本集中的壳向量，然后针对壳向量集进行 SVM 训练，从而获得候选支持向量集。在每次增量学习的过程中，将原壳向量集和新增样本集一起用于 SVM 训练，得到新的分类函数。由于壳向量集是支持向量集的子集，因此，训练中其不但不会丢失支持向量，而且可以比支持向量集更好地代表训练样本集，更多地包含分类信息。文献[93]在壳向量的基础上，利用中心距离比值法选择类边界壳向量进行增量训练，提出了一种基于类边界壳向量的快速 SVM 增量算法(称为 HC-SVM 算法)。该算法可以使训练集的规模显著缩小，从而使得 SVM 的训练速度明显加快，但由于需要通过设定阈值选择边界壳向量，阈值的选择会影响分类器的性能，且阈值的确定方法还没有定论，因此只能通过反复试验来确定。文献[94]将壳向量与 KKT 条件结合对文献[93]进行了改进，提出了基于 KKT 条件与壳向量的增量学习算法。该算法不需要通过主观设定阈值选择边界样本，其首先选择包含所有支持向量的壳向量，再利用 KKT 条件去除新增样本中的冗余样本。实验结果表明，该算法不仅能够保证 SVM 的分类精度和泛化能力，而且学习速度比经典的 SVM 算法更快。

基于增量学习的约简算法利用迭代式思想考虑新增数据对历史数据的影响，若新增数据对分类超平面没有影响，则将其去除，仅保留对分类超平面起作用的数据作为候选支持向量集，这大大缩减了原始数据的规模，同时避免了“过拟合”现象的发生。例如，Mitra 等[89]在 Chinese webpage 数据集上的实验发现，随着增量学习过程的不断进行，算法的准确率由初始的 90.8%上升至 94.3%。但是对于大规模训练集，迭代式增量学习算法将不可避免地会增加算法的计算时间和存储空间，从而影响算法的学习效率。

3.4 基于主动学习的约简算法

主动学习[95]是一个通用术语，它描述了一种特殊的交互式迭代学习过程，可用于使用少量标记数据构建高性能分类器。主动学习与被动学习不同，在被动学习中，学习算法具有一组静态的标记样本，这些样本随后被用于构建模型，而主动学习通过选择信息量最大的样本用于训练。主动学习广泛应用于存在大量未标记数据的情况。主动学习通过一组参数 (H,Q,S,T,U) 表示。其中 H 是一个有监督的分类器；Q 是查询函数，用于查找信息量最大的数据；S 是监督器，负责将真实类别标签分配给未标记的样本；T 是标记样本；U 是未标记样本。首先将有标签的训练集 T 用于在分类器 H 上训练，其次使用查询函数 Q 从未标记样本 U 中选择最具分类信息的数据，并使用监督器 S 为其分配真实的类别标签，新的标记样本将被放入标记样本 T 中，并使用更新的标记样本 T 重新训练分类器 H。如此重复，直到满足某些预定义的停止条件[96]。在主动学习过程中需要定义一些参数，如标记样本 T 和未标记样本 U 的数量。良好的主动学习算法应该对未标记样本 U 的数量不敏感[97]，它应该始终保持良好的性能，而不考虑未标记样本的数量。

考虑到 SVM 分类中边界样本包含最多的分类信息，利用主动学习选择信息量大的样本，不仅可以降低计算复杂度，而且能够避免大规模冗余数据的存储[98-99]。文献[100]指出文献[98]在采样时没有考虑到样本分布的问题，因此其结合

未标记样本 U 的分布使用 K 均值聚类算法对其进行划分,提出了一种具有代表性采样(Representative Sampling)的主动学习算法。该算法优于大部分传统的 SVM 主动学习算法,但是在边缘未标记样本 U 的聚类结构和复杂度较高的情况下,该算法的性能较差。查询函数 Q 是主动学习算法的关键,不同的查询函数对应着不同的主动采样策略。Li 等[101]提出了基于置信度的主动学习算法,该算法根据样本的条件误差来衡量每个样本的置信度值,用于选择未标记样本 U 中不确定性的样本。该算法具有实现简单、计算效率高和鲁棒性强等优点。文献[102]结合壳向量,提出了一种基于凸包向量和样本距离的 SVM 主动学习算法。该算法通过样本距离和壳向量,主动选择对当前 SVM 分类器最有价值的样本,降低了学习中的样本标记成本,增强了 SVM 的泛化性能,提高了训练速度。但该算法仅限于低维数据集,对数据噪声和非线性可分数据敏感,特别不适用于数据分布非凸的情形。文献[103]将主动学习与 K 近邻算法相结合,利用 KNN 算法评估每个样本的稀疏性,将稀疏区域的样本视为信息样本进行选择。但该算法的前提是样本的 K 邻域需包含正负两种样本,若只包含一类样本,则该算法的效果不佳,因此样本的分布和 K 的大小对算法有着显著的影响。文献[104]提出了一种基于决策有向无环图的 SVM 主动学习算法(DDAGB-MASVM 算法),在保持 SVM 性能不变的情况下,使用尽可能少的标记样本训练多类 SVM,或者尽可能提高 SVM 分类的泛化性能。针对 SVM 多分类问题中二分类提供的样本信息可能与多分类不一致的问题,Goudjil 等[105]提出了一种基于后验概率的 SVM 主动学习算法(Multi-classifier AL-SVM,AL-MSVM),该算法使用后验概率模型计算每组数据的平均概率,然后选择那些平均概率低于给定阈值的样本。实验结果表明,所提出的 AL-MSVM 算法提高了分类精度,并保持了数据的稳定性。

基于主动学习的约简算法考虑到 SVM 分类中边界样本包含最多的分类信息,利用查询函数选择信息量大的边界样本作为候选支持向量集。同时,主动学习可通过少量未标记样本构建高性能分类器,这对于大规模标记困难的数据集无疑是至关重要的。与此同时,基于主动学习的约简算法还可通过改变查询函数灵活地选择约简算法,例如基于置信度和 K 近邻的主动学习算法等。在基于置信度的主动学习约简算法中,Li 等[101]在 adult 数据集上的实验结果表明,随着标记训练样本的增加,算法的准确率由原来的 89% 提高到 92%。基于主动学习的约简算法需要设置的参数较多,并且在高维数据上适用性较差,泛化能力较弱。

3.5 基于随机抽样的约简算法

随机抽样是通过随机选择整个样本集的一个子集,并使用该子集代表整个样本集用于 SVM 训练的方法。Lee 等[106]提出的简化支持向量机(Reduced Support Vector Machines,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,但 RSVM 在测试集上的结果甚至比在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。Chang 等[107]将随机抽样技术与增量学习过程相结合,提出了系统采样 RSVM 算法(Systematic Sampling RSVM,SSRSVM)。SSRSVM 从一个非常小的初始约减集开始,并基于当前分类器迭代地将一部分误分类点添加到约减集中,直到验证集正确性达到用户预先指定的阈值。事实证明,SSRSVM 可能会自动生成比随机抽样更小的约减集。此外,在相同水平的测试集正确率下,SSRSVM 比 RSVM 的速度更快,且比传统的 SVM 快得多。文献[108]提出了一种新的方法来生成 RSVM 的代表约减集。首先,引入了聚类简化支持向量机(Clustering Reduced Support Vector Machine,CRSVM),它通过径向基函数(Radial Basis Function,RBF)构建 RSVM 模型。将聚类算法应用于每个类,可以生成每个类的聚类质心,并使用它们形成用于 RSVM 的约减集。此外,算法通过估计每个聚类的近似密度,以获得高斯核中使用的参数,这将节省大量的调优时间(CRSVM 可以为约减集中的每个样本自动确定不同的内核宽度参数)。文献[109]根据分配给训练集的权重(权重越大,该样本为支持向量的可能性越大)随机选择训练子集,使用该训练子集训练 SVM 分类器。再将该 SVM 分类器用于在整个训练集上训练,将分类错误的样本的权重加倍,如果反复迭代,且迭代的次数足够多,训练集中支持向量的权重将比其他向量的权重更大。虽然该方法可以找到一个约简集,但由于初始训练子集的大小事先是未知的,通常以耗时的试错方式确定,因此对于大型数据集,该方法还可能忽略训练子集的分布特征,从而影响 SVM 分类器的性能。因此,如何确定初始训练子集的大小和分布是一个值得研究的问题。

基于随机抽样的约简算法的思想是随机抽取一部分训练集,并将其作为约简后的训练集用于训练。该算法思想简单且容易理解,并且可以显著减小原始训练集的规模。例如,Lee 等[106]在 adult 数据集上的实验结果表明,RSVM 算法至少能够减少 90% 的原始训练。除此之外,Chang 等[107]在 mushroom 数据集上的实验结果表明,与 RSVM 算法相比,SSRSVM 算法训练时间减少了近 50%。尽管训练数据大量减少,但是算法的性能反而得到了显著提升,这也从另一方面说明基于随机抽样的约简算法减少了数据的过拟合。但是基于随机抽样的约简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的效果不理想。

3.6 其他方法

除了以上 5 类 SVM 样本约简算法,学者们还将群智能算法与 SVM 结合,用于寻找最优候选支持向量集。例如,Kawulok 等[110]使用遗传算法从大规模含噪声的数据集中为 SVM 选择有价值的训练数据,提出了基于遗传算法的 GASVM 样本选择算法。该算法能够有效地解决随机抽样依赖于初始数据分布的问题,在 ECU 皮肤数据库上的实验结果表明,相比典型的随机抽样 RSVM 算法,GASVM 算法能够找到分类效果更好的候选支持向量集,并且分类准确率比 RSVM 提高了 1%左右。但是该 GASVM 算法需要手动设置的参数较多,不能自适应调整各个参数。为解决这一问题,Nalepa 等[111]提出了基于自适应遗传算法的 SVM 样本选择算法 AGASVM。该算法引入一种个体间距度量,它能够根据特征空间的样本分布快速分析,并根据当前特征空间的

搜索状态动态调整各个参数。在 adult 和 mushroom 数据集上的实验结果表明，与 GASVM 算法相比，AGASVM 算法的收敛速度更快，并且获得的支持向量数量更少。除此之外，Sharif 等[112]提出了基于多种子的 SVM 样本选择算法 MSB-SVM。该算法适用于处理大规模数据，并且它通过构建最小生成树来处理多类问题，只需要扫描整个数据集一次，算法的时间复杂度为 $O(t\log t)$。在 german 和 liver-disorders 数据集上的实验结果显示，MSB-SVM 的算法准确率分别保持在 79%和 77%左右。

总的来说，不同的约简算法适用范围有所不同，需根据数据集的具体情况选择与之相适的样本约简算法。

3.7 各类 SVM 样本约简算法比较

SVM 样本约简是一种处理大规模训练集行之有效的方法，不仅能够节约存储资源、加快训练速度，并且可以在一定程度上减少过拟合现象的发生，提高分类器的泛化能力。根据不同的样本约简策略，产生了不同种类的约简算法。各类 SVM 样本约简算法性能的详细对比如表 1 所列。此外，为更加清晰地对比各类约简算法，我们从每类算法中挑选了一个代表性算法进行对比，对比结果如表 2 所列。其中，k^+ 和 k^- 分别为正类和负类样本的聚类数量，SV 为候选支持向量集，$m=\frac{1}{\delta}$，$0<\delta\leqslant 1$，δ 为上界因子，t 为样本数量。

表 1 SVM 样本约简算法性能比较

Table 1 Performance comparison of SVM sample reduction algorithms

SVM 样本约简算法	机制	优点	缺点	适用范围	典型算法
基于聚类的约简算法	运用不同的聚类方法去除冗余的样本	快速收敛，具有一定的抗噪能力	对参数和初始样本的设置敏感	大规模低维数据集	SVM-KM[43] FCM[49] CB-SVM[53] DBSCAN[61] MEB[67]
基于几何分析的约简算法	通过构造凸包选择边界样本	算法计算速度快，存储空间小	容易受噪声和离群点的影响	大规模数据集	RCH[73] CCH-SVM[79] CBCH[81]
基于增量学习的约简算法	将数据分批次地进行 SVM 训练，每次只保留支持向量	极大地缩小了原训练集的规模	分批次训练对计算机的存储需求很高，计算成本高	中小型数据集	MR-ISVM[88] HD-SVM[92]
基于主动学习的约简算法	通过查询函数选择信息量最大的样本进行训练	降低了计算复杂度，避免了大规模冗余数据的存储	查询函数的设计没有统一性，对参数取值敏感	少量标记的大型低维数据集	DDAGB-MASVM[104] AL-MSVM[105]
基于随机抽样的约简算法	通过随机选择样本用于 SVM 训练	算法原理简单且容易实现，减少了数据的过拟合	算法迭代成本高，初始训练子集的大小和分布容易影响算法的泛化性能	大规模分布均匀数据集	RSVM[106] SSRSVM[107] CRSVM[108]

表 2 不同代表性算法的性能比较

Table 2 Performance comparison of different representative algorithms

算法	是否为迭代式算法	是否具有抗噪能力	是否会降低分类精度	时间复杂度
WKM-SVM[44]	否	是	否	$O([k^++k^-]^3)+O([(k^++k^-)+\sum_{j\in SV}t_j]^3)$
RCH[73]	是	否	是	$O(t^{m+1}/(m-1)!)$
HD-SVM[92]	是	是	是	—
AL-MSVM[105]	否	否	是	—
RSVM[106]	否	否	否	$O(\bar{t}^2)$，其中 $\bar{t}\leqslant\frac{t}{10}$

3.8 未来与展望

尽管 SVM 样本约简算法得到了广泛的应用，但仍存在一些问题和挑战。结合 SVM 样本约简的研究现状，未来的研究可以从以下 3 方面进行改进：

1)由于数字化和信息化发展日新月异，不断有大量的数据以数据流的形式产生。数据流具有高速到达、动态变化等特点，如何在数据流场景下进行增量学习，已经成为大数据分析领域的研究热点[113]。基于增量学习的 SVM 样本约简算法需要分批次存储数据，对计算机的存储需求和计算能力要求很高。为减少训练时间和计算成本，使用基于数据流的增量学习算法对大规模数据进行样本约简是值得研究的重要问题。与此同时，对大规模标记困难的数据集进行标记也成为了一项艰巨的任务。为此，可考虑将主动学习与半监督学习相结合，并将其应用于半监督学习过程中的样本约简。这也是一个值得考虑的研究方向。

2)初始训练子集对 SVM 分类器至关重要。基于增量学习、主动学习以及随机抽样算法的方法大多需要确立初始训练子集的大小，而现有的算法大多通过随机选择确定，无法根据样本的特征适当调整选择策略。为此，可以发挥基于几何分析约简算法能够考虑样本分布的优势，同时，为避免基于几何分析的约简算法受噪声和离群值影响，可通过设计相关的隶属度函数给不同的样本赋予不同的权重，对噪声和离群值赋予较低的隶属度值，并在样本约简过程中将其约简。将基于随机抽样的约简算法与基于几何分析的约简算法和设计相关的隶属度函数相结合，充分考虑各算法之间的优势和不足，得到的分类器性能可能更佳。

3)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。受其启发,考虑使用模糊隶属度函数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。

结束语 随着大数据时代的快速发展,样本约简是解决 SVM 在处理大规模数据问题上的一种重要方法。尽管 SVM 在小规模数据集上泛化能力很强,但在大规模数据集上仍有很多可探索的空间。本文主要从基于聚类、几何分析、增量学习、主动学习和随机抽样 5 个典型的方向全面概述了 SVM 样本约简学习算法,并讨论了各种 SVM 样本约简算法的优缺点,最后对 SVM 样本约简算法未来的研究提出了展望,希望为新的实践者和理论者提供有价值的参考,以寻求创新方法及应用。

参 考 文 献

[1] RATHORE M M,SHAH S A,SHUKLA D,et al. The role of AI,machine learning,and big data in digital twinning:A systematic literature review,challenges,and opportunities[J]. IEEE Access,2021,9:32030-32052.

[2] DEEPA N,PHAM Q V,NGUYEN D C,et al. A survey on blockchain for big data:Approaches,opportunities,and future directions[J]. Future Generation Computer Systems,2022,131(C):209-226.

[3] TENG K D,ZHAO Q,TAN H R,et al. Emotional EEG recognition based on SVM-KNN algorithm [J]. Computer System Application,2022,31(2):298-304.

[4] GROTHER P J,CANDELA G T,BLUE J L. Fast implementations of nearest neighbor classifiers[J]. Pattern Recognition,1997,30(3):459-465.

[5] OUGIARIGLOU S,DIAMANTARAS K I,EVANGELIDIS G. Exploring the effect of data reduction on neural network and support vector machine classification [J]. Neurocomputing,2018,280:101-110.

[6] CHARBUTY B,ABDULAZEEZ A. Classification based on decision tree algorithm for machine learning[J]. Journal of Applied Science and Technology Trends,2021,2(1):20-28.

[7] LEE C S,CHEANG P Y S,MOSLEHPOUR M. Predictive analytics in business analytics:Decision tree[J]. Advances in Decision Sciences,2022,26(1):1-29.

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

[9] BALASUBRAMANIAM V. Artificial intelligence algorithm with SVM classification using dermascopic images for melanoma diagnosis[J]. Journal of Artificial Intelligence and Capsule Networks,2021,3(1):34-42.

[10] NAEEM M,JAMAL T,DIAZ-MARTINEZ J,et al. Trends and futureperspective challenges in big data[M]// Advances in intelligent data analysis and applications. Singapore:Springer,2022:309-325.

[11] XU X,LIANG T,ZHU J,et al. Review of classical dimensionality reduction and sample selection methods for large-scale data processing[J]. Neurocomputing,2019,328:5-15.

[12] ZHOU Z H. Machine learning[M]. Singapore:Springer Nature,2021:2-17.

[13] YANG J F,QIAO P R,LI Y M,et al. A review of research on machine learning classification problems and algorithms [J]. Statistics and Decision,2019,35(6):36-40.

[14] VAPNIK V,IZMAILOV R. Reinforced SVM method and memorization mechanisms [J]. Pattern Recognition,2021,119:108018.

[15] SZELISKI R. Computer vision:Algorithms and applications [M]. Berlin:Springer Nature,2022:273-300.

[16] MOCHIDA K,KODA S,INOUE K,et al. Computer vision-based phenotyping for improvement of plant productivity:A machine learning perspective[J]. GigaScience,2018,8(1):1-12.

[17] DHAR A,MUKHERJEE H,DASH N S,et al. Text categorization:past and present[J]. Artificial Intelligence Review,2021,54(4):3007-3054.

[18] LUO X. Efficient English text classification using selected machine learning techniques[J]. Alexandria Engineering Journal,2021,60(3):3401-3409.

[19] MOHAMMADI M,RASHID T A,KARIM S H T,et al. A comprehensive survey and taxonomyof the SVM-based intrusion detection systems[J]. Journal of Network and Computer Applications,2021,178:102983.

[20] ZHANG X,LI C,WANG X,et al. A novel fault diagnosis procedure based on improved symplectic geometry mode decomposition and optimized SVM[J]. Measurement,2021,173:108644.

[21] WANG M,CHEN Y,ZHANG X,et al. Roller bearing fault diagnosis based on integrated fault feature and SVM[J]. Journal of Vibration Engineering & Technologies,2022,10(3):853-862.

[22] ALI W,TIAN W,DIN S U,et al. Classical and modern face recognition approaches:A complete review[J]. Multimedia Tools and Applications,2021,80(3):4825-4880.

[23] SALAMH A B S,AKYÜZ H I. A new deep learning model for face recognition and registration in distance learning[J]. International Journal of Emerging Technologies in Learning,2022,17(12):29.

[24] ALWAJIDI S,YANG L. Multiresolution hierarchical support vector machine for classification of large datasets[J]. Knowledge and Information Systems,2022,64(12):1-16.

[25] GOYAL S. Effectivesoftware defect prediction using support vector machines[J]. International Journal of System Assurance Engineering and Management,2022,13(2):681-696.

[26] TANVEER M,RAJANI T,RASTOGI R,et al. Comprehensive review on twin support vector machines[J]. arXiv:2105.00336,2022.

[27] DASH R,DASH D K,PANDA R S. Linguistic information for decision-making using SVM[M]// Advances in data science and management. Singapore:Springer,2022:3-10.

[28] WANG T H,CHEN J T. A review of research on the selection of kernel functions[J]. Computer Engineering and Design,2012,33(3):1181-1186.

[29] ALSHUDUKHI J S. Smart and interactive healthcare system based on speech recognition using soft margin formulation and kernel trick[J]. International Journal of System Assurance Engineering and Management,2022,15:324-333.

[30] WANG Q. Support vector machine algorithm inmachine learning

[C]// 2022 IEEE International Conference on Artificial Intelligence and Computer Applications. IEEE, 2022: 750-756.
[31] 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]. Journal of Biosystems Engineering, 2022, 47(3): 302-317.
[32] NAGPAL M, KAUSHAL M, SHARMA A. A feature reduced intrusion detection system with optimized SVM using big bang big crunch optimization[J]. Wireless Personal Communications, 2022, 122(2): 1939-1965.
[33] HASSANAT A B, ALI H N, TARAWNEHA S, et al. Magnetic force classifier: A novel method for big data classification[J]. IEEE Access, 2022, 10: 12592-12606.
[34] GAYE B, ZHANG D, WULAMU A. Improvement of support vector machine algorithm in big data background[J]. Mathematical Problems in Engineering, 2021, 2021: 1-9.
[35] ALI A A A, MALLAIAH S. Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout[J]. Journal of King Saud University-Computerand Information Sciences, 2022, 34(6): 3294-3300.
[36] KUDO T, MATSUMOTO Y. Chunking with support vector machines[C]// Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics on Language technologies. Association for Computational Linguistics, 2001: 1-8.
[37] SINGH K R, NEETHU K P, MADHUREKAA K, et al. Parallel SVM model for forest fire prediction[J]. Soft Computing Letters, 2021, 3: 100014.
[38] BADR E, ALMOTAIRI S, SALAM M A, et al. New sequential and parallel supportvector machine with grey wolf optimizer for breast cancer diagnosis[J]. Alexandria Engineering Journal, 2022, 61(3): 2520-2534.
[39] DESHPANDE N J, KARIBASAPPA K G, TOTAD S G. Comparative analysis of optimization techniques on multi-classSVM [C] // International Conference for Emerging Technology. IEEE, 2022: 1-6.
[40] QIN Y, DING S F. A review of semi-supervised clustering[J]. Computer Science, 2019, 46(9): 15-21.
[41] HUANG H, XIONG W, WU K et al. K-means hybrid iterative clustering based on memory transfer flagfish optimization [J]. Journal of Shanghai Jiaotong University, 2022, 56(12): 1638.
[42] YAO Y, LIU Y, YU Y, et al. K-SVM: An effective SVM algorithm based on k-means clustering[J]. Journal of Computers, 2013, 8(10): 2632-2639.
[43] ALMEIDA M, BRAGA A, BRAGA J P. SVM-KM: Speeding SVMs learning with a priori cluster selection and k-means[C]// Proceedings of the 6th Brazilian Symposium on Neural Networks. IEEE, 2000: 162-167.
[44] BANG S, JHUN M. Weightedsupport vector machine using k-means clustering[J]. Communications in Statistics—Simulation and Computation, 2014, 43(10): 2307-2324.
[45] KOGGALAGE R, HALGAMUGE S. Reducing the number of training samples for fast support vectormachine classification [J]. Neural Information Processing-Letters and Reviews, 2004, 2(3): 57-65.
[46] SHEN X J, MU L, LI Z, et al. Large-scale support vector machine classification with redundant data reduction[J]. Neurocomputing, 2016, 172: 189-197.
[47] JAVADI S, HASHEMY SHAHDANY S M, NESHAT A, et al. Multi-parameter risk mapping of qazvin aquifer by classic and fuzzy clustering techniques[J]. Geocarto International, 2022, 37(4): 1160-1182.
[48] CERVANTES J, LI X, YU W. Support vector machine classification based on fuzzy clustering for large datasets[C]// Mexican International Conference on Artificial Intelligence. Springer, 2006: 572-582.
[49] MANIMALA K, DAVID I G, SELVI K. A novel dataselection technique using fuzzy c-means clustering to enhance SVM-based power quality classification[J]. Soft Computing, 2015, 19(11): 3123-3144.
[50] ALMASI O N, ROUHANI M. Fast and de-noise support vector machine training method based on fuzzy clustering method for large real world datasets[J]. Turkish Journal of Electrical Engineering and Computer Sciences, 2016, 24(1): 219-233.
[51] WANG L, SUI M, LI Q, et al. A new method of sample reduction for support vector classification[C]// Asia-Pacific Services Computing Conference. IEEE, 2012: 301-304.
[52] KRISHNA D P, SENGUTTUVAN A, LATHA T S. Clustering on large numeric data sets using hierarchical approach: Birch [J]. Global Journal of Computer Science and Technology, 2012, 12(C12): 29-32.
[53] YU H, YANG J, HAN J, et al. Making SVMs scalable to large data sets using hierarchical cluster indexing[J]. Data Mining and Knowledge Discovery, 2005, 11(3): 295-321.
[54] CHU Z, WANG W, LI B, et al. An operation health status monitoring algorithm of special transformers based on birch and Gaussian cloud methods[J]. Energy Reports, 2021, 7: 253-260.
[55] LANG A, SCHUBERT E. Betula: Numerically stable CF-trees forbirch clustering[C]// International Conference on Similarity Search and Applications. Springer, 2020: 281-296.
[56] ALZU'BI A, BARHAM M. Automatic birch thresholding with features transformation for hierarchical breast cancer clustering [J]. International Journal of Electrical and Computer Engineering, 2022, 12(2): 1498-1507.
[57] AWAD M, KHAN L, BASTANI F, et al. An effective support vector machines performance using hierarchical clustering[C]// Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. IEEE, 2004: 663-667.
[58] LUO F, KHAN L, BASTANI F, et al. A dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles [J]. Bioinformatics, 2004, 20 (16): 2605-2617.
[59] HE D J, WU X R, YU L. Research ondensity clustering methods [J]. Communications Technology, 2022, 55(2): 135-142.
[60] WU F F, ZHAO Y L, JIANG Z F. Support vector machine classification algorithm based ondensity clustering [J]. Academic Journal of Xi'an Jiaotong University, 2005, 39(12): 1319-1322.
[61] ZHANG H, ZOU K Q, CUI J, et al. An improved fuzzy support vector machine based on density clustering[J]. Computer Engi-

neering,2009,35(5):194-196.

[62] AKASAPU A K,RAO P S,SHARMA L K,et al. Density based k-nearest neighbors clustering algorithm for trajectory data[J]. International Journal of Advanced Science and Technology,2011,31(1):47-57.

[63] ZHAO F J,HE X S,WANG J. An improved support vector machine based on density clustering[J]. Journal of Jiamusi University:Natural Science Edition,2010,28(4):587-589.

[64] SCHIILKOP P B,BURGEST C,VAPNIK V. Extracting support data for a given task[C]// Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining. AAAI Press,1995:252-257.

[65] SHI P,KUANG L,TANG Y,et al. Pond water quality data stream anomaly detection based on improved SVDD algorithm [J]. Journal of Agricultural Engineering,2021,37(24):249-256.

[66] ZHANG Y,CHI Z X,XIE F D. Improved PCM-based support vector description of multiclass classifier[J]. Computer Science,2008(8):149-153.

[67] CERVANTES J,LI X,YU W,et al. Supportvector machine classification for large data sets via minimum enclosing ball clustering[J]. Neurocomputing,2008,71(4/5/6):611-619.

[68] TSANG I W,KWOK J T,CHEUNG P M,et al. Core vector machines:Fast SVM training on very large data sets[J]. Journal of Machine Learning Research,2005,6(4):363-392.

[69] TSANG I W H,KWOK J T Y,ZURADA J M. Generalized core vector machines[J]. IEEE Transactions on Neural Networks,2006,17(5):1126-1140.

[70] WANG D,ZHANG B,ZHANG P,et al. An online core vector machine with adaptiveMEB adjustment [J]. Pattern Recognition,2010,43(10):3468-3482.

[71] LI D H,DU S X,WU T J. A fast incremental learning algorithm for linear support vector machines based on shell vectors [J]. Journal of Zhejiang University: Engineering Science, 2006,40(2):203-207.

[72] BENNETT K P,BREDENSTEINER E J. Duality and geometry in SVM classifiers[C]// Proceedings of the International Conference on Machine Learning. Citeseer,2000:57-64.

[73] 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,2009:230-239.

[74] CRISP D,BURGES C J. A geometric interpretation of v-SVM classifiers[J]. Advances in Neural Information Processing Systems,1999,12:244-250.

[75] MAVROFORAKIS M E,SDRALIS M,THEODORIDIS SA. Geometric nearest point algorithm for the efficient solution of the SVM classification task[J]. IEEE Transactions on Neural Networks,2007,18(5):1545-1549.

[76] MAVROFORAKIS M E,SDRALIS M,THEODORIDIS S. A novel SVM geometric algorithm based on reduced convex hulls [C]// Proceedings of the 18th International Conference on Pattern Recognition. IEEE,2006:564-568.

[77] MAVROFORAKIS M E,THEODORIDIS S. A geometric approach to support vector machine classification[J]. IEEE Transactions on Neural Networks,2006,17(3):671-682.

[78] LIU Z,LIU J G,PAN C,et al. A novel geometric approach to binary classification based on scaled convex hulls [J]. IEEE Transactions on Neural Networks,2009,20(7):1215-1220.

[79] CHAU A L,LI X,YU W. Large data sets classification using convex-concave hull and support vector machine[J]. Soft Computing,2013,17(5):793-804.

[80] LOPEZ-CHAU A,LI X,YU W. Convex-concave hullfor classification with support vector machine[C]// Proceedings of the 12th International Conference on Data Mining Workshops. IEEE,2012:431-438.

[81] BIRZHANDI P,YOUN H Y. CBCH(clustering-based convex hull) for reducing training time of support vector machine[J]. The Journal of Supercomputing,2019,75(8):5261-5279.

[82] BARBER C B,DOBKIN D P,HUHDANPAA H. The quickhull algorithm for convex hulls[J]. Association for Computing Machinery,1996,22(4):469-483.

[83] WANG D,QIAO H,ZHAN B,et al. Online support vectormachine based on convex hull vertices selection[J]. IEEE Transactions on Neural Networks and Learning Systems,2013,24(4):593-609.

[84] FRANC V,HLAVÁČ V. An iterative algorithm learning the maximal margin classifier[J]. Pattern Recognition,2003,36(9):1985-1996.

[85] KHOSRAVANI H R,RUANO A E,FERREIRA P M. A convex hull-based data selection method for data driven models[J]. Applied Soft Computing,2016,47:515-533.

[86] OSUNA E,CASTRO O D. Convex hull in feature space for support vector machines[C]// Ibero-American Conference on Artificial Intelligence. Springer,2002:411-419.

[87] SYED N A,LIU H,SUNG K K. Incremental learning with support vector machines[C]// Proceedings of the 2001 IEEE International Conference on Data Mining. IEEE,2001:641-642.

[88] XU J,XU C,ZOU B,et al. New incrementallearning algorithm with support vector machines[J]. IEEE Transactions on Systems,Man,and Cybernetics:Systems,2018,49(11):2230-2241.

[89] MITRA P,MURTHY C A,PAL S K. Data condensation in large databases by incremental learning with support vector machines[C]// Proceedings of the 15th International Conference on Pattern Recognition. IEEE,2000:708-711.

[90] CAUWENBERGHS G,POGGIO T. Incremental and decremental support vector machine learning[C]// Proceedings of the 13th International Conference on Neural Information Processing Systems. MIT Press,2000:388-394.

[91] KATAGIRI S,ABE S. Incremental training of support vector machines using hyperspheres[J]. Pattern Recognition Letters,2006,27(13):1495-1507.

[92] LI C,LIU K,WANG H. The incremental learning algorithm with support vector machine based on hyperplane-distance[J]. Applied Intelligence,2011,34(1):19-27.

[93] WU C M,WANG X D,BAI D Y,et al. Fast SVM incremental learning algorithm based on class boundary hull vectors[J]. Computer Engineering and Applications,2010,46(23):185-187.

[94] WEN B,SHAN G L,DUAN X S. Research on incremental learning algorithm based on KKT condition and hull vector [J].

Computer Science, 2013, 40(3): 4.
[95] REN P, XIAO Y, CHANG X, et al. A survey ofdeep active learning[J]. ACM Computing Surveys, 2021, 54(9): 1-40.
[96] DEMIR B, PERSELLO C, BRUZZONE L. Batch-mode active-learning methods for the interactive classification of remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 49(3): 1014-1031.
[97] SASSANO M. An empirical study of active learning with support vector machines for Japanese word segmentation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Natural Language Processing. 2002: 505-512.
[98] SCHOHN G, COHN D. Less is more: active learning with support vector machines[C]// Proceedings of the 17th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2000: 839-846.
[99] TONG S, KOLLER D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2(1): 999-1006.
[100] XU Z, YU K, TRESP V, et al. Representative sampling for text classification using support vector machines [C] // European Conference on Information Retrieval. Springer, 2003: 393-407.
[101] LI M, SETHI I K. Confidence-based active learning [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(8): 1251-1261.
[102] XU H, LI L, GUO P, et al. Uncertainty SVM active learning algorithm based on convex hull and sample distance[C]// Proceedings of the 33th Chinese Control and Decision Conference. IEEE, 2021: 6815-6822.
[103] LENG Y, QI G H, XU X Y, et al. A new SVM active learning algorithm based on KNN[C]// Advanced Materials Research. Trans Tech Publications Ltd., 2014: 2906-2909.
[104] XU H, BIE X, FENG H, et al. Multiclass SVM active learning algorithm based on decision directed acyclic graph and one versus one[J]. Cluster Computing, 2019, 22(3): 6241-6251.
[105] GOUDJIL M, KOUDIL M, BEDDA M, et al. A novel active learning method using SVM for text classification[J]. International Journal of Automation and Computing, 2018, 15(3): 290-298.
[106] 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.
[107] CHANG C C, LEE Y J. Generating the reduced set by systematic sampling[C]// International Conference on Intelligent Data Engineering and Automated Learning. Springer, 2004: 720-725.
[108] CHIEN L J, CHANG C C, LEE Y J. Variant methods of reduced set selection for reduced support vector machines[J]. Journal of Information Science and Engineering, 2010, 26(1): 183-196.
[109] BALCÁZAR J, DAI Y, WATANABE O. A random sampling technique for training support vectormachines [C] // International Conference on Algorithmic Learning Theory. Springer, 2001: 119-134.
[110] KAWULOK M, NALEPA J. Support vector machines training data selection using a genetic algorithm[C]// Proceedings of the 2012 Joint IAPR International Conference on Structural, Syntactic, and Statistical Pattern Recognition. 2012: 557-565.
[111] NALEPA J, KAWULOK M. Adaptive Genetic Algorithm to Select Training Data for Support Vector Machines[C]// Proceedings of the 17th European Conference onApplications of Evolutionary Computation. Springer, 2014: 514-525.
[112] SHARIF I, CHAUDHURI D. A multiseed-based SVM classification technique for training sample reduction[J]. Turkish Journal of Electrical Engineering and Computer Sciences, 2019, 27(1): 595-604.
[113] SUÁREZ C A L, QUINTANA D, CERVANTES A. A survey on machine learning for recurring concept drifting data streams [J]. Expert Systems with Applications, 2022, 213(2): 118934.

ZHANG Daili, born in 1996, postgraduate. Her main research interests include machine learning and data mining.

WANG Tinghua, born in 1977, Ph. D, professor, is a senior member of CCF (No. 95071S). His main research interests include artificial intelligence and machine learning.

(责任编辑:何杨)