文章编号:2096-1472(2025)10-0017-05
DOI:10.19644/j.cnki.issn2096-1472.2025.010.004

基于聚类表征的支持向量机分类方法

蒋露露,沈项军

(江苏大学计算机科学与通信工程学院,江苏 镇江 212013)
✉ 2544998926@qq.com; xjshen@ujs.edu.cn

摘 要:传统的支持向量机分类算法在面对分类环境中的复杂噪声问题时往往存在局限性。针对这一问题提出一种基于聚类表征的支持向量机分类方法。利用样本之间的距离信息自适应学习具有一定概率邻域分配的结构化图。此外,通过邻域样本构造一个新的数据样本能够增强其数据表征,有助于更好地避免分类器受到噪声干扰。通过原数据下以及大量噪声诱导场景下的各项实验验证,实验结果平均提升了4%,证明该方法在解决分类噪声问题上具有显著的效果。

关键词:支持向量机;样本空间;噪声;聚类;表征学习
中图分类号:TP181 **文献标志码**:A

Support Vector Machine Classification Method Based on Clustering Representation

JIANG Lulu, SHEN Xiangjun

(*School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang* 212013, *China*)
✉ 2544998926@qq.com; xjshen@ujs.edu.cn

Abstract: Traditional Support Vector Machine (SVM) classification algorithms often exhibit limitations when dealing with complex noise problems in classification environments. To address this issue, this paper proposes a clustering representation-based SVM classification method. It adaptively learns a structured graph with probabilistic neighborhood assignments by utilizing distance information between samples. Furthermore, constructing a new data sample using neighborhood samples enhances its data representation, which helps the classifier better avoid noise interference. Experimental results conducted on the original data and under numerous noise-induced scenarios demonstrate that the proposed method achieves an average improvement of 4 percentage points, proving its significant effectiveness in solving classification noise problems.

Key words: support vector machine (SVM); sample space; noise; clustering; representation learning

0 引言(Introduction)

支持向量机(Support Vector Machine,SVM)作为一种监督学习算法,致力于寻找一个最优的超平面,以最大化不同类别样本之间的间隔[1-2]。然而,与大多数监督学习算法一样,SVM基于训练集中的标记数据学习分类器,而没有特定的策略来处理噪声数据,从而导致分类性能欠佳。

与监督学习不同,无监督学习可以深入内部揭示样本空间的底层结构,揭示样本之间的内在关系[3]。聚类作为一种无监督学习任务,根据样本的距离或相似性将相似的样本分组在一起,其目标在于捕捉样本及其邻域之间的相关性。

受上述内容启发,本文提出了一种将监督学习和无监督学习结合起来的混合模型,以应对噪声问题给模型分类器带来的

收稿日期:2024-12-09
基金项目:国家自然科学基金项目(62376108)

干扰。具体而言，在传统的软间隔支持向量机模型基础上，引入结构化图聚类模型来增强输入样本邻域之间的特征相关性。此外，构建一个新的数据样本，使得模型可以更好地利用样本之间的相似性信息，并对支持向量机的决策面进行调整，以更好地区分不同的类别。

1 相关工作(Related work)

1.1 支持向量机

支持向量机(SVM)是在统计学习理论的基础上发展起来的一种机器学习算法，也是目前应用最广泛的分类算法之一。经典 SVM 分类器最早是被 CORTES 等[1]提出的。这个分类器是由样本构建并依赖一个带有 Mercer 核的再生核希尔伯特空间，能够处理高维数据集并且具有很好的泛化能力。然而，在一些复杂模型中往往存在着噪声和异常值的干扰，这会弱化 SVM 的分类效果。为了解决这个问题，许多新方法应运而生。WANG 等[4]提出了一种广义非凸损失函数的鲁棒支持向量机，通过牛顿算法求解，具有更好的鲁棒性能和回归性能。LI 等[5]利用噪声使 SVM 的分离裕度成为随机变量，提出了一种概率支持向量机(PSVM)来捕获分离裕度的概率信息，并在这种噪声环境中制定决策函数。LIN 等[6]提出了模糊支持向量机(FSVM)分类方法，该方法通过使用噪声或异常值对数据进行分类，从而反方向减少了噪声或异常值对于分类任务的影响。然而，这些方法都是让分类模型更适应于噪声环境，考虑到的问题都是如何从含有噪声和异常值的环境中获得更优的分类效果。与上述方法不同的是，本文提出的方法从样本源头一定程度上消除噪声和异常值，提升原有数据表征能力，以达到遏制其对于分类效果的影响。

1.2 聚类

作为一种无监督学习任务，聚类旨在将数据集中相似的样本归为一类，并将不同的样本归为不同的类别[7-8]。其中，基于图的聚类通过利用数据之间的局部相似性构建图结构从而捕捉数据之间的内在关系，取得了显著的性能[9]。基于图的聚类性能的关键在于构建能够准确捕捉样本内在关系的高质量相似图。为此，CHENG 等[10]利用同一子空间数据对应的稀疏表示来构建相似图。ZHU 等[11]将局部邻域样本关系编码为具有从所有可用特征中识别出判别特征的相似图。然而，它们构造的图不能保证具有理想的聚类结构，而聚类结构对聚类性能很重要。于是，文献[12-13]中的方法被开发出来，通过学习具有更好内部结构的相似图来保证具有理想的聚类结构。然而，真实数据中通常含有噪声和离群条目，可能会破坏数据相似图的可靠性，从而降低聚类性能。本文引入的聚类结构在自适应学习具有适当概率邻域分配的结构化图的同时进行聚类[14]，能够有效地抵抗噪声和异常值。

2 本文方法(Proposed method)

本节将详细讨论提出的基于聚类表征的支持向量机模型。模型主要分为 3 个阶段：首先，阐述了聚类方法，该方法的目标是学习一个稳健的特征表示；其次，将聚类方法引入软间隔支持向量机模型，从而将学习到的稳健的特征表示作为分类器的输入；最后，在本文的非凸学习框架中使用增强拉格朗日乘法(ALM)来优化求解相关参数。

2.1 结构化图聚类学习

聚类方法目标是学习一个鲁棒的特征表示作为结构化图学习的输入，使图能够抵抗噪声，过滤异常值，并保持原始数据的局部结构。给定一个数据矩阵 $\boldsymbol{X}\in\mathbb{R}^{n\times d}$，其中 n 为样本总数，d 为特征维数。在原始特征空间中引入一个亲和度矩阵 $\boldsymbol{S}\in\mathbb{R}^{n\times n}$ 来度量数据中每对样本点之间的相似程度。聚类结构的优化目标如式(1)所示：

$$\begin{aligned}&\sum_{i=1}^{N}\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\boldsymbol{x}_{i,j}\\&\text{s.t.}\,\text{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1\end{aligned}\tag{1}$$

其中：$\boldsymbol{S}_{i,j}$ 为 $\boldsymbol{x}_i$ 连接到所有样本的概率，即第 i 个数据点到第 j 个数据点的距离权值。为了能够更好地捕捉到相似数据点之间的关系，在为每个数据点分配适当的邻居时，较大的权值对应着较小的距离，较小的权值则反之。约束中的 $\text{diag}(\boldsymbol{S})=0$ 意味着每个数据点与自身之间不具有亲和关系；$\boldsymbol{S}^{\mathrm{T}}1=1$ 表示每个数据点在整个数据集中的亲和度权重和为 1。

2.2 基于聚类表征的支持向量机分类方法

传统的支持向量机通过在原始数据特征上最大化数据样本的边缘来获得分类器，导致鲁棒性差，在噪声场景下泛化能力弱。因此在本文提出的方法中，将上述提到的聚类方法和软间隔支持向量机集成到统一框架中，这能够增强原始数据样本之间特征的相关性，更有效地获取样本数据中最具代表性的特征信息，从而使得模型在噪声环境下拥有更高的分类性能。具体来说，定义一个数据集 $T=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$，其中，$\boldsymbol{x}_i\in\boldsymbol{X}\in\mathbb{R}^{n\times d}$，$\boldsymbol{x}_i$ 是第 i 个特征向量，y_i 是特征 $\boldsymbol{x}_i$ 的类别标签。得到目标函数如式(2)所示：

$$\begin{aligned}&\min_{w,S}\frac{\lambda_1}{2}\|\boldsymbol{w}\|_2^2+C\sum_i^N\xi_i+\frac{\lambda_2}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\boldsymbol{S}_{i,j}+\\&\frac{\lambda_3}{2}\|\boldsymbol{S}\|_F^2\\&\text{s.t.}\,y_i\left(\boldsymbol{w}^{\mathrm{T}}\left(\boldsymbol{x}_i+\lambda_4\sum_j^N\boldsymbol{S}_{i,j}\boldsymbol{x}_j\right)+b\right)\geqslant 1-\xi_i,\forall i,\xi_i>0,\\&\text{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1\end{aligned}\tag{2}$$

其中：λ_1、λ_2、λ_3、λ_4 为正则化参数，C 是惩罚参数，最后一项为正则化以避免平凡解。通过对邻域样本加权构造了一个新的数据样本来增强原始数据的表示，本质上是一个拉普拉斯平滑过程.将增强后的特征表示作为支持向量机分类器的输入，有效地避免了噪声对分类效果的干扰。为了简化表达式，引入 $\widetilde{\boldsymbol{x}}_i=\boldsymbol{x}_i+\lambda_3\sum_j^N\boldsymbol{S}_{i,j}\boldsymbol{x}_j$，得到目标函数及其改进的约束形式如式(3)所示：

$$\begin{aligned}&\min_{w,\boldsymbol{S}}\frac{\lambda_1}{2}\|\boldsymbol{w}\|_2^2+C\sum_i^N\xi_i+\frac{\lambda_2}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\boldsymbol{S}_{i,j}+\\&\frac{\lambda_3}{2}\|\boldsymbol{S}\|_F^2\\&\text{s.t.}\,y_i(\boldsymbol{w}^{\mathrm{T}}\widetilde{\boldsymbol{x}}_i+b)\geqslant 1-\xi_i,\forall i,\xi_i>0,\text{diag}(\boldsymbol{S})=0,\\&\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1\end{aligned}\tag{3}$$

2.3 优化求解

由于提出的模型涉及了正则化优化问题，传统的优化算法可能无法直接处理这些约束，这里引入 ALM 优化方法来解决

这一问题。将原始优化问题转化为增广拉格朗日函数的形式,引入拉格朗日乘子,并使用惩罚项对约束进行建模。这种转化将非凸约束表示为凸函数,再通过迭代的方式进行求解。

对于式(3),引入拉格朗日乘子 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_n)$,$\beta=(\beta_1,\beta_2,\cdots,\beta_n)$,引用 Karush-Kuhn-Tucker 定理[15]作为约束条件来推导对偶问题。优化问题的拉格朗日函数为

$$L(\boldsymbol{\xi},\alpha,\beta,\boldsymbol{w},\boldsymbol{S})=\frac{\lambda_1}{2}\|\boldsymbol{w}\|_2^2+C\sum_i^N\xi_i+\frac{\lambda_2}{2}\sum_{i=1}^N\sum_{j=1}^N\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\boldsymbol{S}_{i,j}+\frac{\lambda_3}{2}\|\boldsymbol{S}\|_F^2+\sum_i^N\alpha_i(1-\xi_i-y_i(\boldsymbol{w}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i+b))-\sum_i^N\beta_i\xi_i$$

$$\text{s.t.}\ \forall i,\xi_i>0,\operatorname{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1,\tilde{\boldsymbol{x}}_i=\boldsymbol{x}_i+\lambda_4\sum_j^N\boldsymbol{S}_{i,j}\boldsymbol{x}_j \tag{4}$$

优化目标是通过最小化合页损失函数和正则化项,找到最优的权重向量 $\boldsymbol{w}$ 和偏置 b。通常采用梯度下降法更新参数,通过式(5)~式(7)获取最优解。

$$\frac{\partial L}{\partial w}=0\rightarrow\boldsymbol{w}=\frac{1}{\lambda_1}\sum_i^N\alpha_i y_i\tilde{\boldsymbol{x}}_i \tag{5}$$

$$\frac{\partial L}{\partial b}=0\rightarrow\sum_i^N\alpha_i y_i=0 \tag{6}$$

$$\frac{\partial L}{\partial\boldsymbol{\xi}}=0\rightarrow\beta_i=C-\alpha_i \tag{7}$$

使用对偶变量代替方程得到目标函数如式(8)所示:

$$\max_{\alpha}\min_{\boldsymbol{S}}L(\boldsymbol{\alpha},\boldsymbol{S})=\frac{\lambda_2}{2}\sum_{i=1}^N\sum_{j=1}^N\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\boldsymbol{S}_{i,j}+\frac{\lambda_3}{2}\|\boldsymbol{S}\|_F^2-\frac{\lambda_1}{2}\sum_i^N\sum_j^N\alpha_i\alpha_j y_i y_j\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\tilde{\boldsymbol{x}}_j+\sum_i^N\alpha_i$$

$$\text{s.t.}\operatorname{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1,\tilde{\boldsymbol{x}}_i=\boldsymbol{x}_i+\lambda_4\sum_j^N\boldsymbol{S}_{i,j}\boldsymbol{x}_j \tag{8}$$

此时需要解决模型中的约束问题,这里本文用到 ALM 优化方法将非凸约束近似为凸函数,并引入辅助变量 $\boldsymbol{H}=\boldsymbol{S}$ 解耦,得到模型相应完整的增广拉格朗日函数如式(9)所示:

$$L(\alpha,\boldsymbol{S},\boldsymbol{H},\boldsymbol{M}_1,\boldsymbol{M}_2)=\frac{\lambda_2}{2}\sum_{i=1}^N\sum_{j=1}^N\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\boldsymbol{S}_{i,j}+\frac{\lambda_3}{2}\|\boldsymbol{S}\|_F^2-\frac{\lambda_1}{2}\sum_i^N\sum_j^N\alpha_i\alpha_j y_i y_j\tilde{\boldsymbol{x}}_i^{\mathrm{T}}\tilde{\boldsymbol{x}}_j+\sum_i^N\alpha_i+\frac{u}{2}\|\boldsymbol{H}-\boldsymbol{S}+\frac{\boldsymbol{M}_1}{u}\|_F^2+\frac{u}{2}\|\tilde{\boldsymbol{X}}-(\boldsymbol{X}+\lambda_4\boldsymbol{X}\boldsymbol{H})+\frac{\boldsymbol{M}_2}{u}\|_F^2$$

$$\text{s.t.}\operatorname{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1 \tag{9}$$

其中:u 为正则化参数,$\boldsymbol{M}_1\in\mathbb{R}^{n\times n}$,$\boldsymbol{M}_2\in\mathbb{R}^{n\times d}$,是拉格朗日乘子,将约束条件与目标函数进行融合,以达到对约束条件的逼近。

2.3.1　更新参数 $\boldsymbol{S}$

针对 $\boldsymbol{S}$ 变量优化目标函数,同时固定其余变量。迭代步骤详细见式(10):

$$\min_{\boldsymbol{S}}\frac{\lambda_2}{2}\sum_{i=1}^N\sum_{j=1}^N\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\boldsymbol{S}_{i,j}+\frac{\lambda_3}{2}\|\boldsymbol{S}\|_F^2+\frac{u}{2}\|\boldsymbol{H}-\boldsymbol{S}+\frac{\boldsymbol{M}_1}{u}\|_F^2$$

$$\text{s.t.}\operatorname{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1 \tag{10}$$

为了式子表达方便起见,令 $\sum_{i=1}^N\sum_{j=1}^N\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2=\boldsymbol{G}_{i,j}$,得到

$$\min_{\boldsymbol{S}}\frac{\lambda_2}{2}\operatorname{tr}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{S})+\frac{\lambda_3}{2}\|\boldsymbol{S}\|_F^2+\frac{u}{2}\|\boldsymbol{H}-\boldsymbol{S}+\frac{\boldsymbol{M}_1}{u}\|_F^2 \tag{11}$$
$$\text{s.t.}\operatorname{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1$$

为了便于计算,令 $\boldsymbol{H}-\boldsymbol{S}+\frac{\boldsymbol{M}_1}{u}$。注意到式(11)在不同的 i 值之间是独立的,因此对于每个 i 分别处理如式(12)所示问题。

$$\min_{\boldsymbol{S}_i}\|\boldsymbol{S}_i+(\lambda_2\boldsymbol{G}_i-2u\boldsymbol{N}_i)/(u+\lambda_3)\|_2^2 \tag{12}$$
$$\text{s.t.}\operatorname{diag}(\boldsymbol{S})=0,\boldsymbol{S}\geqslant 0,\boldsymbol{S}^{\mathrm{T}}1=1$$

考虑到约束,有

$$\min_{\boldsymbol{S}_i}\sum_{\boldsymbol{S}}\|\boldsymbol{S}_i+(\lambda_2\boldsymbol{G}_i-2u\boldsymbol{N}_i)/(u+\lambda_3)\|_2^2-\eta(1^{\mathrm{T}}\boldsymbol{S}_i-1)-\zeta^{\mathrm{T}}\boldsymbol{S}_i \tag{13}$$
$$\text{s.t.}\operatorname{diag}(\boldsymbol{S})=0$$

对 $\boldsymbol{S}_i$ 求导并令其为 0,有

$$\boldsymbol{S}_i+(\lambda_2\boldsymbol{G}_i-2u\boldsymbol{N}_i)/(u+\lambda_3)-\eta 1-\zeta=0 \tag{14}$$

$\boldsymbol{S}_i$ 的第 j 项如式(15)所示:

$$\boldsymbol{S}_{ij}+(\lambda_2\boldsymbol{G}_{ij}-2u\boldsymbol{N}_{ij})/(u+\lambda_3)-\eta-\zeta_i=0 \tag{15}$$

根据 KKT 条件,得到最终优化结果:

$$\boldsymbol{S}_{ij}=(\eta-(\lambda_2\boldsymbol{G}_{ij}-2u\boldsymbol{N}_{ij})/(u+\lambda_3))_+ \tag{16}$$

2.3.2　更新参数 $\boldsymbol{H}$

针对 $\boldsymbol{H}$ 变量优化目标函数,同时固定其余变量。迭代步骤为

$$\min_{\boldsymbol{H}}\frac{u}{2}\|\boldsymbol{H}-\boldsymbol{S}+\frac{\boldsymbol{M}_1}{u}\|_F^2+\frac{u}{2}\|\tilde{\boldsymbol{X}}-(\boldsymbol{X}+\lambda_4\boldsymbol{X}\boldsymbol{H})+\frac{\boldsymbol{M}_2}{u}\|_F^2 \tag{17}$$

将其表示为含有 H 的拉格朗日表达式,并设为 0,求导得

$$\boldsymbol{H}=(1+\lambda_4^2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\left(\boldsymbol{S}-\frac{\boldsymbol{M}_1}{u}+\lambda_4\boldsymbol{X}^{\mathrm{T}}\frac{\boldsymbol{M}_2}{u}+\lambda_4\boldsymbol{X}^{\mathrm{T}}\tilde{\boldsymbol{X}}-\lambda_4\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right) \tag{18}$$

算法 1 总结了基于聚类表征的支持向量机分类算法的详细求解步骤。

算法 1　基于聚类表征的支持向量机分类算法

输入:训练数据 $\boldsymbol{X}\in\mathbb{R}^{m\times n}$,参数 $\lambda_1,\lambda_2,\lambda_3,\lambda_4$

初始化:$\boldsymbol{M}_1=\boldsymbol{M}_2=0,\boldsymbol{w}=0,\boldsymbol{S}=0,\boldsymbol{H}=0$

迭代求解以下步骤直至损失值达到收敛条件或者达到最大迭代次数:

步骤 1:固定其余变量,通过式(5)更新 $\boldsymbol{w}$。

步骤 2:固定其余变量,通过式(16)更新 $\boldsymbol{S}$。

步骤 3:固定其余变量,通过式(18)更新 $\boldsymbol{H}$。

输出:$\boldsymbol{w}$。

3 实验与分析(Experiment and analysis)

本节将在多个流行的基准数据集上展示提出的方法以及其他几个先进的支持向量机分类算法的性能比较,以评估本文所提方法的有效性和稳健性。为了使得实验结果更具说服力,将从不同的实验角度进行实验验证。所有比较方法中的参数都经过仔细调整,以报告最佳性能。

3.1 实验数据集

本文选取的4个大规模数据集及其基本信息见表1。

表1 数据集

Tab.1 Data set

数据名称	样本数/个	特征维度
MNIST	14 780	784
USPS	19 999	784
COIL10	720	1024
Batch	10 000	3072

(1)MNIST:1998年11月由纽约大学的库兰特研究所、谷歌纽约实验室和微软雷德蒙德研究院发布的手写数字数据集。该数据集共包含14 780个28×28像素的灰度样本。

(2)USPS:美国邮政服务自动扫描信件获得的数字识别数据集。该数据集包含19 999个28×28像素的灰度样本。

(3)COIL10:由不同物体图像组成的数据集,其中包含10个物体,720个样本。

(4)Batch:一个图像数据集,总共包含10 000个图像样本。

3.2 评估基线

实验中用于比较的最先进的分类算法有:

(1)SVM[1]:传统的软间隔支持向量机模型。

(2)CMLS[16]:该算法通过将来自不同模式的神经成像数据投射到潜在空间中来工作。随后,它将获得的潜在表征转换到标签空间中以训练不同的分类器。最后,采用集成方法提高分类结果的可靠性。

(3)MSVM[17]:基于最大化多类边距的多类支持向量机。

(4)CappedSVM[18]:利用一个新的基于上限 l_p 范数的目标,提出了一种新的上限 l_p 支持向量机分类模型。

3.3 评价指标

为了评估各种算法的分类结果,本研究采用准确率作为评价指标对结果进行分析。每个数据集随机选择70%的实例作为训练集,其余的分配给测试集。此外,为了证明结果的可靠性,每种测试方法将在每个数据集上执行10次,随后,将推导平均值以确定最终结果。本实验采用精度作为评价指标,计算过程如式(19)所示:

$$A=\frac{TN+TP}{FN+FP+TN+TP} \tag{19}$$

其中:FN 表示原始数据 i 类中被分类为 j 类的样本数量;同样地,FP 表示把原本未被分类为 i 类的样本分类为 i 类的数量;TN 表示原本不属于 i 类的也未被分类为 i 类的样本数量;TP 表示被正确分类为 i 类的 i 类样本数量。

3.4 实验结果

在本节中,首先在4个大规模数据集上展示了提出方法与评估基线的比较结果,并给出了其中一个数据集上各方法的混淆矩阵。然后,分别验证了所提方法以及评估基线在不同程度的椒盐噪声以及高斯噪声诱导下的数据集上的性能,从而证明本文方法在噪声场景下的有效性。

3.4.1 原数据下的对比结果

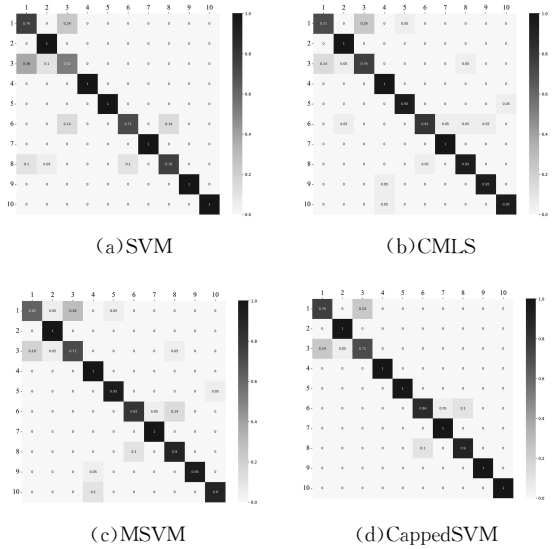
表2给出了本文方法以及其他4种分类对比算法在不同图像数据集上的平均准确率。可以看出,本文方法在4个数据集上的性能均优于其他对比算法,在COIL10数据集上的准确率最高,达到93.46%,相比SVM、CMLS、MSVM、CappedSVM算法分别高出5.41%、2.20%、5.26%和1.15%。由于Batch数据集的维数和样本量比较大,数据集更为复杂,几个算法的平均准确率也会相对较低一点,但本文提出的方法结果相比其他算法结果仍然可以获得比较大的提升,可以看出本文提出方法在复杂数据集中的表现效果好,甚至分类效果有更显著的提升。

表2 各算法在图像数据集上的平均准确率

Tab.2 Average accuracy of each algorithm on image data set

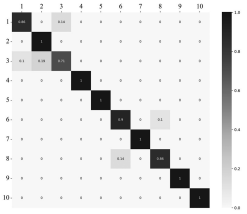
数据集	平均准确率/%				
	SVM	CMLS	MSVM	CappedSVM	本文方法
MNIST	85.26	90.33	87.21	90.97	**92.07**
USPS	85.03	90.78	88.04	91.08	**92.15**
COIL10	88.05	91.26	88.20	92.31	**93.46**
Batch	80.16	82.28	80.96	83.77	**85.82**

为了进一步直观地评价提出模型的分类性能,图1展示了本文提出方法和各种对比方法在COIL10数据集上的混淆矩阵。其中,混淆矩阵由表示预测类别的列和表示实际类别的行组成。对角线值表示在分类过程中做出的准确预测的比例。此外,非对角线元素表示预测的部分是错误的,并且色块的颜色越深,比例越高。相比其他4种分类算法,提出方法的混淆矩阵对角线上的色块颜色整体最深,其中,数值为1的色块个数最多,且其他不为0的色块数量最少。这说明提出方法的正确分类的类别及样本数量最多,错误分类的类别及样本数量最少。这验证了提出方法的分类准确率高,分类效果好。



(a)SVM (b)CMLS

(c)MSVM (d)CappedSVM



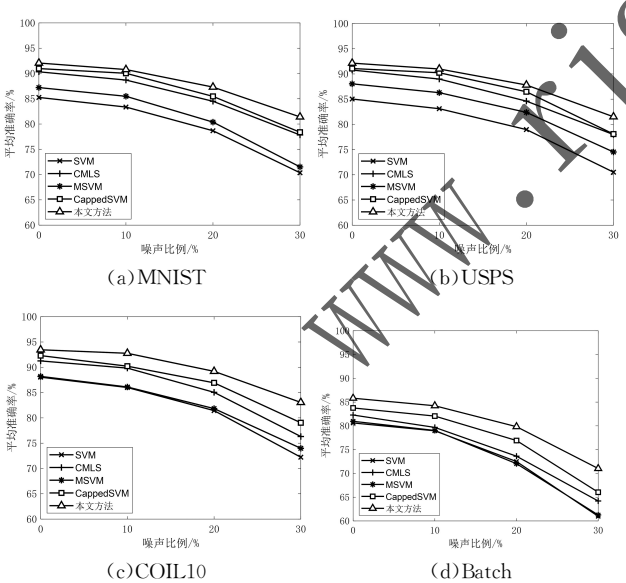
(e)本文方法

图 1　各种算法在 COIL10 数据集上的混淆矩阵

Fig. 1 Confusion matrix of algorithms on COIL10 data set

3.4.2　不同程度噪声数据下的对比结果

如图 2 所示,比较了 SVM、CMLS、MSVM、CappedSVM 以及本文方法在不同程度椒盐噪声影响的 MNIST、USPS、COIL10 和 Batch 数据集上的性能。其中所加入的椒盐噪声的比例分别为 10%、20%和 30%。在 4 个不同的数据集上,所提方法在加入 30%椒盐噪声后的精度相比原数据集下降了平均 11.52%,分别比相同情况下 SVM、CMLS、MSVM、CappedSVM 算法低了 4.70%、3.06%、4.27%、2.63%。可以看出在有椒盐噪声的环境下,本文算法的性能表现更优越。并且随着噪声水平的提高,本文算法曲线的下降趋势相比其他算法更为平缓,说明提出方法的有效性越来越明显。这一点在 Batch 数据集上更为明显。由于 Batch 数据集上的样本数及维度较高,平均准确率及下降趋势较其他数据集更为明显,但其他对比算法相比提出方法的差距也更为明显,可以说明本文方法在复杂数据集中的性能优势更为显著。



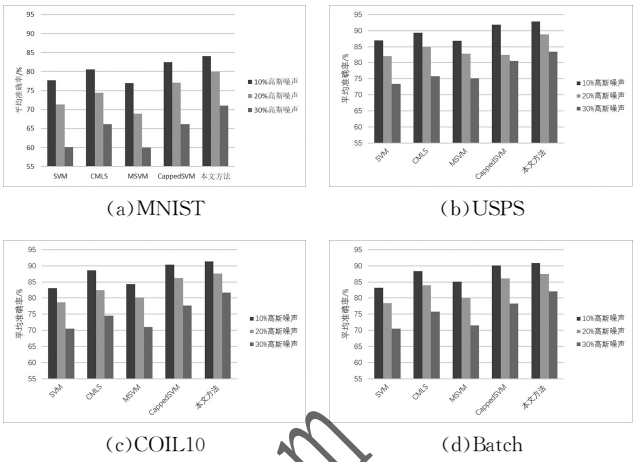
(a)MNIST　(b)USPS

(c)COIL10　(d)Batch

图 2　不同椒盐噪声比例影响下不同算法的平均准确率

Fig. 2 Average accuracy of different algorithms under the influence of different proportions of salt and pepper noise

图 3 显示了 MNIST、USPS、COIL10 和 Batch 数据集上高斯噪声相对于平均准确率评估指标性能的相应变化。所加入的椒盐噪声的比例分别为 10%、20%和 30%,可以观察到其性能变化趋势类似于在椒盐噪声场景下的趋势变化。在 4 个不同的数据集上,本文所提方法的平均准确率在 10%、20%和 30%的高斯噪声影响下均高于相同场景下的其他算法。并且随着噪声的增加,准确率下降趋势更为平缓,说明本文提出方法在一定程度上的抗噪效果更好,在噪声场景下的鲁棒性和泛化能力更好。



(a)MNIST　(b)USPS

(c)COIL10　(d)Batch

图 3　不同程度高斯噪声影响下的不同算法平均准确率

Fig. 3 Average accuracy of different algorithms under the influence of different degrees of Gaussian noise

4　结论(Conclusion)

本文提出了一种基于聚类表征的支持向量机分类方法,以应对复杂图像分类环境中的噪声和异常值干扰问题。该方法创新性地将聚类结构和支持向量机模型结合到统一框架中,通过聚类表征学习来增强分类模型的特征表示,大大提高了分类器在高噪声条件下的分类性能。通过一系列原始条件以及更多噪声加持条件下的实验结果表明,本文所提分类方法在噪声场景下的分类性能明显优于其他相关分类方法。在未来,计划引入深度学习方法来进一步增强分类模型。

参考文献(References)

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

[2] MELACCI S, BELKIN M. Laplacian support vector machines trained in the primal[J]. Journal of machine learning research, 2011,12:1149-1184.

[3] HUANG Y, KECHADI T. An effective hybrid learning system for telecommunication churn prediction[J]. Expert systems with applications, 2013,40(14):5635-5647.

[4] WANG K, ZHU W, ZHONG P. Robust support vector regression with generalized Loss function and applications [J]. Neural processing letters, 2015,41(1):89-106.

[5] LI H, YANG J, ZHANG G, et al. Probabilistic support vector machines for classification of noise affected data[J]. Information sciences, 2013, 221(2):60-71.

[6] LIN C F, WANG S D. Training algorithms for fuzzy support vector machines with noisy data[J]. Pattern recognition letters, 2004, 25(14):1647-1656.

(下转第 46 页)

[7] BONEH D, LYNN B, SHACHAM H. Short signatures from the weil pairing[M]. International conference on the theory and application of cryptology and information security. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001:514-532.
[8] 崔彬. 云数据完整性验证技术综述[J]. 信息技术与信息化, 2020(12):8-10.

作者简介:

邵 琪(1999-),男(汉族),徐州,硕士生。研究领域:智能信息处理。
刘从军(1974-),男(汉族),武汉,高级实验师,硕士。研究领域:智能信息处理。本文通信作者。

(上接第 21 页)

[7] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1):48-61.
[8] HAN D Y, KIM J. Unified Simultaneous Clustering and Feature Selection for Unlabeled and Labeled Data[J]. IEEE transactions on neural networks and learning systems, 2018,29(12):1-16.
[9] WU Z, LEAHY R. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 1993, 15(11):1101-1113.
[10] CHENG B, YANG J C, YAN S C, et al. Learning with l1-graph for image analysis[J]. IEEE transactions on image processing: a publication of the IEEE signal processing society, 2010, 19(4):858-866.
[11] ZHU X T, LOY C C, GONG S G. Constructing Robust Affinity Graphs for Spectral Clustering[C]// IEEE. IEEE Conference on Computer Vision & Pattern Recognition. Columbus: IEEE, 2014:1450-1457.
[12] NIE F P, WANG X Q, HUANG H. Clustering and projected clustering with adaptive neighbors [M]. ACM, 2014:977-986.
[13] NIE F P, WANG X Q, JORDAN M I, et al. The Constrained laplacian rank algorithm for graph-Basedclustering[C] // AAAI Press. Proceedings of the AAAI conference on artificial intelligence. Phoenix: AAAI Press. 2016, 30(1):86-102.
[14] SHI D, ZHU L, LI Y K, et al. Robust structured graph clustering[J]. IEEE transactions on neural networks and learning systems, 2019, 31(11):4424-4436.
[15] CRISTIANINI N, SHAWETAYLOR J. An introduction to support vector machines[M]. Beijing, China Machine Press, 2005:16-17.
[16] ZHOU T, THUNG K H, LIU M X, et al. Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging Data[J]. Medical image analysis, 2019, 60:101630.
[17] XU J, LIU X L, HUO Z Y, et al. Multi-Class support vector machine via maximizing multi-class margins[C] // NSF Public Access Repository(PAR). The 26th International Joint Conference on Artificial Intelligence (IJCAI 2017). Eisenhower Avenue: NSF Public Access Repository(PAR). 2017:3154-3160.
[18] NIE F P, WANG X Q, HUANG H. Multiclass capped p-Norm SVM for robust classifications[C] // AAAI. Proceedings of the AAAI conference on artificial intelligence. Phoenix: AAAI. 2017, 31(1):435-450.

作者简介:

蒋露露(2000-),女(汉族),无锡,硕士生。研究领域:图像处理和机器学习。本文通信作者。
沈项军(1977-),男(汉族),镇江,教授,博士。研究领域:模式识别与智能决策系统。

(上接第 30 页)

[10] WANG Z, ZHOU X, ZHAO W, et al. Assessing the severity of sleep apnea syndrome based on ballistocardiogram [J]. PloS one, 2017, 12(4):e0175351.
[11] Al-ANGARI H M, SAHAKIAN A V. Use of sample entropy approach to study heart rate variability in obstructive sleep apnea syndrome[J]. IEEE transactions on biomedical engineering, 2007, 54(10):1900-1904.
[12] CAO K, LV X. Multi-task feature fusion network for obstructive sleep apnea detection using single-lead ECG signal[J]. Measurement, 2022, 202:111787.

作者简介:

李轩逸(1998-),男(汉族),温州,硕士生。研究领域:机器学习与智能应用。
朱耀东(1970-),男(汉族),南通,教授,博士。研究领域:智能机器人与健康物联网。本文通信作者。
赵伟岚(1999-),男(汉族),邵阳,硕士生。研究领域:机器学习与智能应用。
樊紫岩(1998-),女(汉族),邢台,硕士生。研究领域:机器学习与智能应用。