山西大学学报(自然科学版)47(1):40—47,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023133

基于邻域粒化的逻辑回归算法

闫静茹[1],陈颖悦[1*],曾高发[2],刘培谦[1],傅兴宇[3]
(1.厦门理工学院 经济与管理学院,福建 厦门 361024;
2.厦门市执象智能科技有限公司,福建 厦门 361000;
3.厦门理工学院 计算机与信息工程学院,福建 厦门 361024)

摘 要:逻辑回归作为一种经典的分类算法,其结构简单且可解释性强。然而,逻辑回归难以处理模糊与不确定的非线性数据。为了解决这一问题,通过采用粒计算理论中的邻域粒化技术,提出了一种基于邻域粒化的逻辑回归算法。对于非线性数据,邻域粒化使数据更容易进行分离和构造。首先,对数据集样本的单特征进行邻域粒化,构造出邻域粒子。然后在多特征上形成邻域粒向量。此外,定义了这些邻域粒向量的度量与运算规则,并设计了一种邻域粒逻辑回归算法,有效地提高了逻辑回归的分类准确性。在WDBC(Diagnostic Wisconsin Breast Cancer),Iris以及Seeds等数据集上进行了分类实验,与经典的逻辑回归进行了比较,结果表明,本文提出算法的分类准确率相较于经典的逻辑回归在三个数据集上分别高出0.6%,7.6%,4.1%。

关键词:逻辑回归;单特征粒化;粒计算;邻域粒子;粒向量
中图分类号:TP181 **文献标志码**:A **文章编号**:0253-2395(2024)01-0040-08

Logistic Regression Algorithm Based on Neighborhood Granulation

YAN Jingru[1], CHEN Yingyue[1*], ZENG Gaofa[2], LIU Peiqian[1], FU Xingyu[3]
(1. *School of Economics and Management, Xiamen University of Technology, Xiamen* 361024*, China;*
2. *Xiamen Zhixiang Intelligent Technology Co., Ltd., Xiamen* 361000*, China;*
3. *College of Computer and Information Engineering, Xiamen University of Technology, Xiamen* 361024*, China*)

Abstract: As a classical classification algorithm, logistic regression has a simple structure and strong interpretability. However, logistic regression is difficult to deal with fuzzy and uncertain nonlinear data. To solve this problem, a logistic regression algorithm based on neighborhood granulation is proposed by using neighborhood granulation technology in granular computing theory. For nonlinear data, neighborhood granulation makes the data easier to separate and construct. Firstly, the neighborhood granules are constructed by the neighborhood granulation of the single feature of the data set sample. The neighborhood granular vectors are then formed on the multi-feature. In addition, the measurement and operation rules of these neighborhood granular vectors are defined, and a neighborhood granular logistic regression algorithm is designed, which effectively improves the classification accuracy of logistic regression. Classification experiments are carried out on WDBC (Diagnostic Wisconsin Breast Cancer), Iris and Seeds data sets, and compared with the classical logistic regression.The results show that the classification accuracy of the proposed algorithm is 0.6%, 7.6% and 4.1% higher than that of the classical logistic regression in the three data sets, respectively.

Key words: logistic regression; single feature granulation; granular computing; neighborhood granules; granular vectors

收稿日期:2023-05-06;**接受日期**:2023-07-24
基金项目:厦门市科技计划项目(2022CXY0428)
作者简介:闫静茹(1997-),女,山西吕梁人,硕士研究生,研究方向为粒计算。E-mail:yjr19121695187@163.com
* **通信作者**:陈颖悦(CHEN Yingyue),E-mail:cyyyjslw@163.com
引文格式:闫静茹,陈颖悦,曾高发,等.基于邻域粒化的逻辑回归算法[J].山西大学学报(自然科学版),2024,47(1):40-47. DOI:10.13451/j.sxu.ns.2023133

0 引言

粒计算是人工智能研究领域的一个方向,涵盖了模糊集、粗糙集与三支决策理论等。1965年,Zadeh[1]提出了模糊集理论,模糊集通过提取模糊信息和不确定信息进行建模解决问题。1982年,Pawlak[2]提出了粗糙集理论,粗糙集也是处理模糊信息和不确定信息的数学工具,信息系统是粗糙集理论的基本概念。1988年,Lin[3]提出了邻域系统,并在1996年首次提出粒计算的概念,粒计算的缩写为GrC。此后,粒计算理论吸引了大量的学者进行研究,并出现了多种粒计算的研究方向。为进一步发展粗糙集理论,众多学者将邻域系统引入粗糙集,提出了邻域粒计算。1998年,Yao[4]提出了一个邻域粗糙集的框架,并引入距离函数来计算邻域。2008年,胡清华等[5]将邻域粗糙集模型引入分类器,提出了一种概念简单且易于实现的方法来理解和构建邻域分类器。同年,胡清华等[6]提出了邻域粒化并用于数值属性约简,改善了分类能力。研究结果表明粒计算能有效应用在各分类器上。2019年,陈玉明等[7]提出了粒向量的概念,并将邻域粒化和K近邻分类器结合,设计有高精度的粒分类器。2022年,张清华等人[8]改进了邻域分类器,有效改善了邻域分类器的泛化能力。邻域粗糙集对粒计算的发展有着巨大的贡献,提供了新的理论基础和应用模型[9-11]。粒计算从不同的粒度空间对复杂问题进行求解,进而提升机器学习的性能[12-14]。

分类问题是机器学习的重要研究内容之一,在分析由特征描述的数据库元组时具有极其重要的作用。分类的目标就是对未知样本进行分类,同一类别中样本的相似度应尽可能大,而不同类别中样本之间的相似度应尽量小。逻辑回归是机器学习中的一种分类算法,国内外大量学者对其进行研究讨论,并在众多领域中广泛被使用。Abramovich等[15]通过稀疏多项式逻辑回归处理高维多分类问题。Sha[16]使用逻辑回归将情绪分为积极情绪和消极情绪,提高了在情绪分析预测中的精度。Castilla等[17]提出具有复杂设计的多项式逻辑回归,使其具备较好的鲁棒性。王敬等[18]提出了一类广义稀疏逻辑回归算法,并成功应用于全脑fMRI数据分类。Pan等[19]引入了一种新的L1正则化逻辑回归安全特征消除规则,该规则允许封闭形式的解决方案,并且不会重复扫描识别的特征。Liu等[20]将粗糙集与逻辑回归算法结合,决策理论粗糙集采用三向决策来避免逻辑回归的错误分类。

粒计算相关的理论比较完善,其在机器学习和深度学习领域的应用广泛,而随着模糊不确定性数据量的与日俱增,传统的逻辑回归难以处理模糊不确定性数据。Diao等[21]将模糊粒化和逻辑回归结合,成功提高个体手势分类的准确性。但是,模糊粒化不能进行参数调整,所提高的分类精度有限。为了处理集合类型的模糊数据和不确定的非线性数据,本文将逻辑回归算法与粒计算理论结合起来,采取邻域粒化技术,提出了基于邻域粒化的逻辑回归算法。对样本做邻域粒化处理,经过粒化后的多维特征数据形成邻域粒向量。通过使用邻域粒向量的度量和运算规则,定义了粒函数与公式。根据粒函数与公式,进一步设计了邻域粒逻辑回归分类算法。最后,在多个UCI(University of California, Irvine)数据集上进行分类实验。实验结果表明,在最优邻域参数下,基于邻域粒化的逻辑回归算法比原始的逻辑回归算法分类效果更优。

本文的其余部分组织如下。第1节介绍了邻域粒化与粒向量。第2节介绍了邻域粒向量的运算。随后,我们提出了基于邻域粒化的逻辑回归算法,解决了逻辑回归模糊数据和不确定的非线性数据问题。在第3节中,进行了一些实验来证明我们提出的方法的有效性。最后,对本文进行总结,并在第4节中进行了一些评论和讨论。

1 邻域粒化与粒向量

粗糙集是处理不确定性信息的有效工具。邻域粒化通常通过邻域粗糙集的二元关系得到,其邻域粒化是在整个邻域粒空间中进行的。下面给出了一系列的定义。

定义1 设分类信息系统为$L=(X,T,B)$,其中样本集表示$X=\{x_1,x_2,\cdots,x_n\}$,特征集表

示 $T=\{t_1,t_2,\cdots,t_k\}$,$B=\{b\}$表示类别。

定义2 设分类信息系统为 $L=(X,T,B)$,对于 $\forall x_i,x_j\in X$ 和单特征 $\forall t\in T$,则 x_i 与 x_j 在单特征 t 上的相似度公式为

$$d_t(x_i,x_j)=1-\left|f_t(x_i)-f_t(x_j)\right|, \quad (1)$$

其中 $d_t(x_i,x_j)\in[0,1]$,$f_t(x_i)$ 和 $f_t(x_j)$ 分别表示 x_i 和 x_j 是在单特征 t 上归一化后的值。

定义3 设分类信息系统为 $L=(X,T,B)$,对于 $\forall x_i,x_j\in X$ 和单特征 $\forall t\in T$,给定一个邻域参数 δ,则样本 x_i,x_j 的邻域判别函数定义为

$$\varphi_t(x_i,x_j)=\begin{cases}0, & d_t(x_i,x_j)>\delta,\\ 1, & d_t(x_i,x_j)\leqslant\delta,\end{cases} \quad (2)$$

其中,当 $\varphi_t(x_i,x_j)=1$,则说明样本 x_i,x_j 相邻;反之,当 $\varphi_t(x_i,x_j)=0$,则说明样本 x_i,x_j 没有相邻。

定义4 设分类信息系统为 $L=(X,T,B)$,对于 $\forall x\in X$ 和单特征 $\forall t\in T$,则 x 在单特征 t 中进行邻域粒化,形成的邻域粒子定义为

$$g_t(x)=\{z_j\}_{j=1}^n=\{z_1,z_2,\cdots,z_n\}, \quad (3)$$

其中 $z_j=\varphi_t(x_i,x_j)$ 为样本 x_i,x_j 在单特征 t 上的邻域判别函数,进行判定两个样本是否互为邻域。

定义5 设分类信息系统为 $L=(X,T,B)$,对于 $\forall x\in X$,任一特征子集 $E\subseteq T$,设 $E=\{t_1,t_2,\cdots,t_k\}$,则 x 在特征子集 E 中的邻域粒向量定义为

$$\boldsymbol{G}_E(x)=(g_1(x),g_2(x),\cdots,g_k(x))^T, \quad (4)$$

其中 $g_k(x)$ 为样本 x 在单特征 t_k 中的邻域粒子。

邻域粒子是一个有序集合,该集合由0或1组成,表示样本之间是否互为邻域。而邻域粒子组成邻域粒向量,所以邻域粒向量也是一个由0或1构成的有序集合。

例 1 存在一个分类信息系统 $L=(X,T,B)$,如表1所示。设一个邻域参数 $\delta=0.5$,一个特征子集 $E=\{t_1,t_2,t_3\}$。

样本集 $X=\{x_1,x_2,x_3,x_4\}$,在单特征 t_1 上进行邻域粒化时,形成邻域粒子分别为

$g_{t_1}(x_1)_{0.5}=\{0,0,0,0\}$,$g_{t_1}(x_2)_{0.5}=\{0,0,1,0\}$,

$g_{t_1}(x_3)_{0.5}=\{0,1,0,0\}$,$g_{t_1}(x_4)_{0.5}=\{0,0,0,0\}$。

样本集 $X=\{x_1,x_2,x_3,x_4\}$,在单特征 t_2 上进行邻域粒化时,形成邻域粒子分别为

表1 分类信息系统 $L=(X,T,B)$

Table 1 The classification information system is $L=(X,T,B)$

X	t_1	t_2	t_3	B
x_1	0.4	0.6	0.1	0
x_2	0.7	0.6	0.3	0
x_3	0.2	0.1	0.9	1
x_4	0.3	0.2	0.8	1

注:其中,$X=\{x_1,x_2,x_3,x_4\}$表示一个样本集,$T=\{t_1,t_2,t_3\}$表示一个特征集,$B=\{0,1\}$表示一个类别集

$g_{t_2}(x_1)_{0.5}=\{0,0,1,0\}$,$g_{t_2}(x_2)_{0.5}=\{0,0,1,0\}$,

$g_{t_2}(x_3)_{0.5}=\{1,1,0,0\}$,$g_{t_2}(x_4)_{0.5}=\{0,0,0,0\}$。

样本集 $X=\{x_1,x_2,x_3,x_4\}$,在单特征 t_3 上进行邻域粒化时,形成邻域粒子分别为

$$g_{t_3}(x_1)_{0.5}=\{0,0,1,1\},$$
$$g_{t_3}(x_2)_{0.5}=\{0,0,1,1\},$$
$$g_{t_3}(x_3)_{0.5}=\{1,1,0,0\},$$
$$g_{t_3}(x_4)_{0.5}=\{1,1,0,0\}。$$

样本 x_1 在 E 中的邻域粒向量为

$$\boldsymbol{G}_E(x_1)_{0.5}=(g_{t_1}(x_1)_{0.5},g_{t_2}(x_1)_{0.5},g_{t_3}(x_1)_{0.5})^T=(\{0,0,0,0\},\{0,0,1,0\},\{0,0,1,1\})^T。$$

样本 x_2 在 E 中的邻域粒向量为

$$\boldsymbol{G}_E(x_2)_{0.5}=(g_{t_1}(x_2)_{0.5},g_{t_2}(x_2)_{0.5},g_{t_3}(x_2)_{0.5})^T=(\{0,0,1,0\},\{0,0,1,0\},\{0,0,1,1\})^T。$$

样本 x_3 在 E 中的邻域粒向量为

$$\boldsymbol{G}_E(x_3)_{0.5}=(g_{t_1}(x_3)_{0.5},g_{t_2}(x_3)_{0.5},g_{t_3}(x_3)_{0.5})^T=(\{0,1,0,0\},\{1,1,0,0\},\{1,1,0,0\})^T。$$

样本 x_4 在 E 中的邻域粒向量为

$$\boldsymbol{G}_E(x_4)_{0.5}=(g_{t_1}(x_4)_{0.5},g_{t_2}(x_4)_{0.5},g_{t_3}(x_4)_{0.5})^T=(\{0,0,0,0\},\{0,0,0,0\},\{1,1,0,0\})^T。$$

2 基于邻域粒化的逻辑回归算法

逻辑回归是一种对数线性模型,在二分类应用广泛,但逻辑回归对多分类的效果却不理想。将粒计算引入逻辑回归,设计粒分类器,以提升在二分类和多分类的性能。在引入粒分类器之前,首先给出邻域粒向量的运算规则。

2.1 邻域粒向量的运算

定义6 设 $g_{t_1}(x)$ 和 $g_{t_2}(x)$ 分别是样本 x 在单特征 t_1 和单特征 t_2 上的两个邻域粒子,则邻域粒子的加法、减法和乘法运算定义如下

$$g_{t_1}(x)+g_{t_2}(x)=\{g_{t_1}(x)_j+g_{t_2}(x)_j\}_{j=1}^n, \quad (5)$$

$$g_{t_1}(x)-g_{t_2}(x)=\{g_{t_1}(x)_j-g_{t_2}(x)_j\}_{j=1}^n, \quad (6)$$

$$g_{t_1}(x)\times g_{t_2}(x)=\{g_{t_1}(x)_j\times g_{t_2}(x)_j\}_{j=1}^{n}。\quad(7)$$

定义7 设两个邻域粒向量分别为

$$\boldsymbol{G}(x_1)=(g_1(x_1),g_2(x_1),\cdots,g_k(x_1))^T,$$

$$\boldsymbol{G}(x_2)=(g_1(x_2),g_2(x_2),\cdots,g_k(x_2))^T,$$

则定义两个邻域粒向量的点积运算为

$$\boldsymbol{G}(x_1)\cdot\boldsymbol{G}(x_2)=g_1(x_1)\times g_1(x_2)+g_2(x_1)\times g_2(x_2)+\cdots+g_k(x_1)\times g_k(x_2)。\quad(8)$$

定义8 设一个邻域粒向量为$\boldsymbol{G}(x)=(g_1(x),g_2(x),\cdots,g_k(x),1)^T$和一个权值邻域粒向量$W=(w_1,w_2,\cdots,w_k,b)^T$,则点积运算为

$$W\cdot\boldsymbol{G}(x)=w_1\times g_1(x)+w_2\times g_2(x)+\cdots+w_k\times g_k(x)+1\times b,\quad(9)$$

两个邻域粒子经过运算后的结果是一个邻域粒子,两个邻域粒向量的点积得到一个邻域粒子,由此来设计粒分类器。

2.2 邻域粒逻辑回归

定义9 设分类信息系统为$L=(X,T,B)$,邻域粒子$g(x)=\{z_j\}_{j=1}^{n}$,$g(X)$是连续随机的粒子,$g(X)$满足粒逻辑分布,则分布粒函数和密度粒函数的定义为

$$F(g(x))=P(g(X)\leqslant g(x))=\left\{\frac{1}{1+\mathrm{e}^{-(z_j-\alpha)/\beta}}\right\}_{j=1}^{n},\quad(10)$$

$$f(g(x))=F'(g(x))=\left\{\frac{\mathrm{e}^{-(z_j-\alpha)/\beta}}{\beta(1+\mathrm{e}^{-(z_j-\alpha)/\beta})^2}\right\}_{j=1}^{n},\quad(11)$$

其中,α为位置粒子参数,$\beta>0$为形状粒子参数。

定义10 设分类信息系统为$L=(X,T,B)$,邻域粒子$g(x)=\{z_j\}_{j=1}^{n}$,标签$B=\{0,1\}$,权值邻域粒向量$W=(w_1,w_2,\cdots,w_k,b)^T$,则邻域粒逻辑回归的条件概率粒分布定义为

$$P(B=1|g(x))=\left\{\frac{\mathrm{e}^{-(w\cdot z_j+b)}}{1+\mathrm{e}^{-(w\cdot z_j+b)}}\right\}_{j=1}^{n},\quad(12)$$

$$P(B=0|g(x))=\left\{\frac{1}{1+\mathrm{e}^{-(w\cdot z_j+b)}}\right\}_{j=1}^{n}。\quad(13)$$

定义11 设分类信息系统为$L=(X,T,B)$,邻域粒子$g(x)=\{z_j\}_{j=1}^{n}$,$P=(B=1|g(x))=\pi(g(x))$,$P=(B=0|g(x))=1-\pi(g(x))$,则邻域粒逻辑回归的似然粒函数和对数似然粒函数定义为

$$\prod_{m=1}^{H}[\pi(g(x)_m)]^{b_m}[1-\pi(g(x)_m)]^{1-b_m},\quad(14)$$

$$L(w)=\left\{\prod_{m=1}^{H}[b_m(w\cdot(z_j)_m)-\log(1+\mathrm{e}^{-(w\cdot(z_j)_m)})]\right\}_{j=1}^{n}。\quad(15)$$

逻辑回归的输出映射在0到1,而邻域粒逻辑回归预测的概率值同样在0到1。具体的基于邻域粒化的逻辑回归算法如算法1所示。

算法1 基于邻域粒化的逻辑回归算法

输入:分类信息系统为$L=(X,T,B)$,邻域参数δ。

输出:概率最大的类别。

1)样本集合X进行邻域粒化为邻域粒向量$GB=(\boldsymbol{G}_T(x_1),\boldsymbol{G}_T(x_2),\cdots,\boldsymbol{G}_T(x_k))^T$;

2)B有n个类别,则建立n个邻域粒逻辑回归模型;

3)每个邻域粒逻辑回归模型以其中一类作为1类,其他作为0类;

4)通过定义11求出n个权值邻域粒向量;

5)通过步骤4求出的权值邻域粒向量代入定义10学习到n个邻域粒逻辑回归模型;

6)将n个学习后的邻域粒逻辑回归模型对邻域粒向量进行预测,得到预测值;

7)预测值归一化,得到属于各个类别的概率;

8)输出概率最大的类别即为预测类别。

3 实验分析

本节通过对比实验验证了所提分类算法的有效性。在邻域粒化中,邻域参数的选择影响分类的准确性。在实验中,首先对本文提出的基于邻域粒化的逻辑回归算法进行了邻域参数的测试。然后将本文提出的基于邻域粒化的逻辑回归算法与几种经典的分类算法进行了比较。

本文从UCI机器学习数据库[22]中下载了6个公共数据集,详情如表2所示。为了确保所有数据的取值范围都能转换为数值在[0,1],使用了最大最小值归一化公式对数据集进行处理

$$\hat{x}_{ab}=\frac{x_{ab}-x_b^{\min}}{x_b^{\max}-x_b^{\min}}。\quad(16)$$

在评估过程中,随机选取70%的数据作为训练集,其余数据作为测试集,验证一次。此操作重复5次,因此计算平均准确率作为性能参数进行评估。

表2 所选数据集的统计数据

Table 2 The statistics of the chosen datasets

数据集名称	样本数目	特征数目	类别数目
WDBC	569	30	2
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Ionosphere	351	34	2
Haberman	306	3	2

3.1 邻域参数的影响

本节将原始的逻辑回归算法作为参照对象对邻域粒化的参数进行了分析,并通过实验结果研究了邻域参数对算法准确率的影响。邻域参数设置为0.05到0.95,间隔为0.05。不同数据集上的分类性能对比如图1—图6所示。

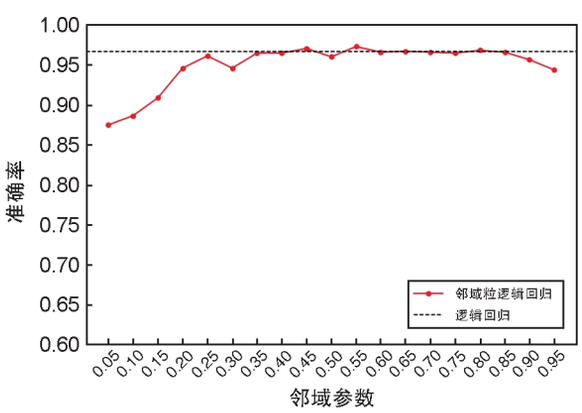

图1 在WDBC数据集上邻域参数对粒逻辑回归准确率的影响

Fig. 1 Effect of neighborhood parameters on accuracy of granular logistic regression on WDBC dataset

由图1可得,对于WDBC(Diagnostic Wisconsin Breast Cancer)数据集,原始的逻辑回归的分类准确率为0.967 3。当邻域参数为0.65时,基于邻域粒化的逻辑回归算法分类准确率等于原始的逻辑回归算法。当邻域参数为0.45、0.55和0.80时,基于邻域粒化的逻辑回归算法分类准确率大于原始的逻辑回归算法。当邻域参数为0.55时,基于邻域粒化的逻辑回归算法分类准确率达到最大值0.967 3。

由图2可得,对于Iris数据集,原始的逻辑回归的分类准确率为0.902 2。邻域参数从0.05到0.30,基于邻域粒化的逻辑回归算法分类准确率不断上升到达最大值0.977 8。邻域参数从0.25到0.95,基于邻域粒化的逻辑回归算法分

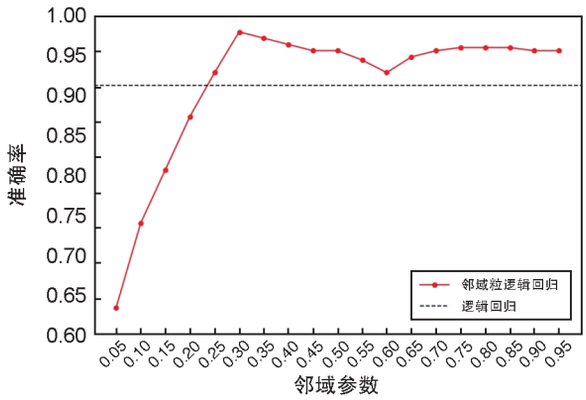

图2 在Iris数据集上邻域参数对粒逻辑回归准确率的影响

Fig. 2 Effect of neighborhood parameters on accuracy of granular logistic regression on Iris dataset

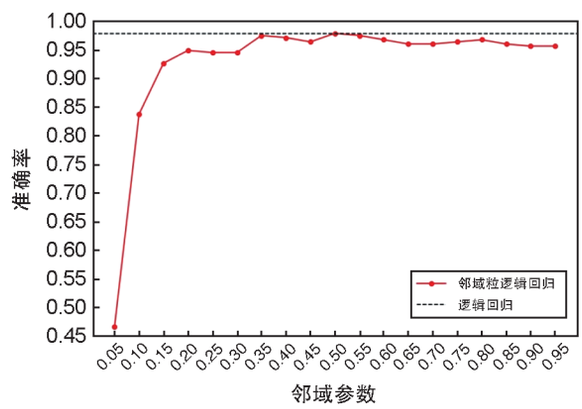

图3 在Wine数据集上邻域参数对粒逻辑回归准确率的影响

Fig. 3 Effect of neighborhood parameters on accuracy of granular logistic regression on Wine dataset

类准确率均大于原始的逻辑回归算法。

由图3可得,对于Wine数据集,原始的逻辑回归的分类准确率为0.977 8。邻域参数只有等于0.50时,基于邻域粒化的逻辑回归算法分类准确率等于原始的逻辑回归算法。邻域参数在0.35之后,基于邻域粒化的逻辑回归算法分类准确率趋于稳定。

由图4可得,对于Seeds数据集,原始的逻辑回归的分类准确率为0.901 6。邻域参数从0.05到0.50,基于邻域粒化的逻辑回归算法分类准确率不断上升到达最大值0.942 9。并且当邻域参数为0.60和0.75时,基于邻域粒化的逻辑回归算法分类准确率也到达最大值。邻域参数从0.35到0.95,基于邻域粒化的逻辑回归算法分类准确率优于原始的逻辑回归算法。

由图5可得,对于Ionosphere数据集,原始的逻辑回归的分类准确率为0.888 7。邻域参数

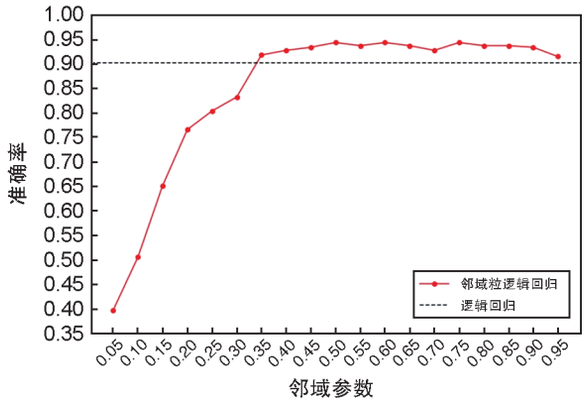

图4 在Seeds数据集上邻域参数对粒逻辑回归准确率的影响

Fig. 4 Effect of neighborhood parameters on accuracy of granular logistic regression on Seeds dataset

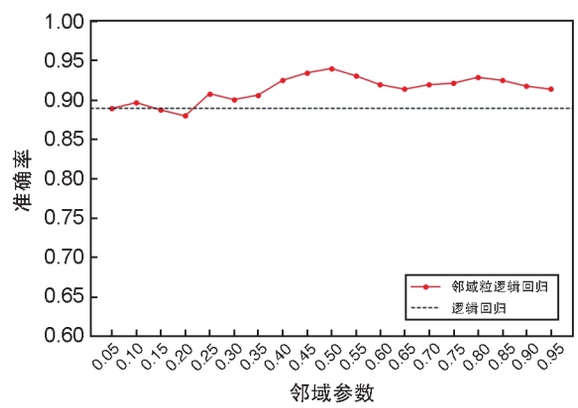

图5 在Ionosphere数据集上邻域参数对粒逻辑回归准确率的影响

Fig. 5 Effect of neighborhood parameters on accuracy of granular logistic regression on Ionosphere dataset

为0.15和0.20时,基于邻域粒化的逻辑回归算法分类准确率略微低于原始的逻辑回归算法;在其他邻域参数上,基于邻域粒化的逻辑回归算法分类准确率均高于原始的逻辑回归算法。当邻域参数为0.50时,基于邻域粒化的逻辑回归算法分类准确率到达最大值0.939 6。

由图6可得,对于Haberman数据集,原始的逻辑回归的分类准确率为0.754 3。当邻域参数等于0.10和0.25时,基于邻域粒化的逻辑回归算法分类准确率等于原始的逻辑回归算法。当邻域参数为0.15、0.20和0.30时,基于邻域粒化的逻辑回归算法分类准确率略微优于原始的逻辑回归算法。当邻域参数为0.30时,基于邻域粒化的逻辑回归算法分类准确率到达最大值0.776 1。

从图1—图6可以看出,对于不同的数据

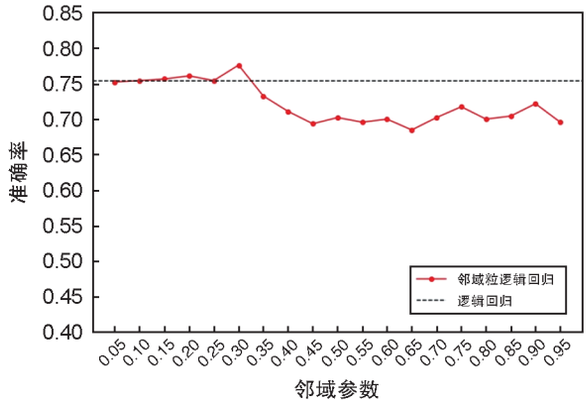

图6 在Haberman数据集上邻域参数对粒逻辑回归准确率的影响

Fig. 6 Effect of neighborhood parameters on accuracy of granular logistic regression on Haberman dataset

集,邻域参数对粒分类器的准确率有着很大的影响。在每个实验数据集中,总能找到合适的邻域参数,这使得基于邻域粒化的逻辑回归算法的分类准确率略微优于原始的逻辑回归算法。

3.2 多种分类算法比较

本节实验主要是将所提出的基于邻域粒化的逻辑回归算法与经典的机器学习分类算法进行比较。经典的分类算法有k近邻、支持向量机、逻辑回归、随机森林和梯度提升决策树。选取3.1节中分类准确率最高的邻域参数作为本节粒逻辑回归的邻域参数。分类算法平均预测精度的具体数值,如表3所示。

从表3可以看出,在WDBC和Haberman数据集上,经典分类算法中支持向量机有着最高的分类精度分别为0.971 9和0.763,而基于邻域粒化的逻辑回归算法的分类精度略优于支持向量机分别为0.973 1和0.776 1。在Iris和Seeds数据集上,经典分类算法中k近邻有着最高的分类精度分别为0.973 3和0.933 3,而基于邻域粒化的逻辑回归算法的分类精度略优于k近邻分别为0.977 8和0.942 9。在Wine数据集上,经典分类算法中原始的逻辑回归有着最高的分类精度0.977 8,基于邻域粒化的逻辑回归算法的分类精度跟原始的逻辑回归相同;但基于邻域粒化的逻辑回归算法的标准差为0.024 1比原始的逻辑回归0.033 1数值更小,更加稳定。在Ionosphere数据集上,经典分类算法中支持向量机有着最高的分类精度0.939 6,基于邻域粒化的逻辑回归算法

表3 粒逻辑回归的平均预测精度(均值±标准差%)和几种经典分类算法

Table 3 The average prediction accuracy (mean±standard deviation %) of granular logistic regression and several classical classification algorithms

数据集	粒逻辑回归	k近邻	支持向量机	逻辑回归	随机森林	梯度提升树
WDBC	97.31±0.89	96.96±0.96	97.19±1.40	96.73±1.58	94.39±2.05	95.20±1.27
Iris	97.78±2.22	97.33±2.43	96.89±2.53	90.22±2.98	96.44±2.53	95.56±2.22
Wine	97.78±2.41	95.93±3.04	97.41±2.11	97.78±3.31	97.41±2.81	91.11±6.73
Seeds	94.29±1.81	93.33±2.35	92.70±2.41	90.16±4.40	91.75±4.26	92.70±3.09
Ionosphere	93.96±1.43	83.58±4.50	93.96±1.71	88.87±3.01	93.77±0.84	92.64±2.05
Haberman	77.61±5.57	70.87±5.56	76.30±5.46	75.43±4.46	72.17±2.25	72.39±2.62

的分类精度跟支持向量机相等;但基于邻域粒化的逻辑回归算法的标准差为0.014 3比支持向量机0.017 1数值更小,更加稳定。

4 结论

本文通过粒计算的思想,利用邻域粒化技术,设计了一种似然粒函数,构建了基于邻域粒化的逻辑回归算法,通过在UCI数据集上进行分类实验,验证了提出的方法性能优于经典的机器学习算法,解决了传统的逻辑回归算法无法对模糊不确定性数据集进行高效地分类。此外,解决了模糊粒化无法调参的问题,使粒逻辑回归算法在适当的邻域参数条件下能取得较好的分类效果。

在未来的工作中,尝试研究新的粒化方法来提高分类性能,并将其应用到实际的生活领域中,如信用卡欺诈、客户流失和疾病预测等。并尝试研究基于粒计算的融合算法,将本文提出的邻域粒逻辑回归和其他算法融合,全面提高粒计算的速度和性能。

参考文献:

[1] ZADEH L A. Fuzzy Sets[J]. *Inf Control*, 1965, **8**(3): 338−353. DOI: 10.1016/s0019-9958(65)90241-x.

[2] PAWLAK Z. Rough Sets[J]. *Inter J Comput Inform Sci*, 1982, **11**(5): 341−356. DOI: 10.1007/BF01001956.

[3] LIN T Y. Neighborhood Systems and Relational Databases[C]//Proceedings of the 1988 ACM Sixteenth Annual Conference on Computer Science. New York: ACM, 1988: 725. DOI: 10.1145/322609.323183.

[4] YAO Y Y. Relational Interpretations of Neighborhood Operators and Rough Set Approximation Operators[J]. *Inf Sci*, 1998, **111**(1/2/3/4): 239−259. DOI: 10.1016/S0020-0255(98)10006-3.

[5] HU Q H, YU D R, XIE Z X. Neighborhood Classifiers [J]. *Expert Syst Appl*, 2008, **34**(2): 866−876. DOI: 10.1016/j.eswa.2006.10.043.

[6] 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简[J]. 软件学报, 2008, **19**(3): 640−649. DOI: 10.3724/SP.J.1001.2008.00640.
HU Q H, YU D R, XIE Z X. Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation[J]. *J Softw*, 2008, **19**(3): 640−649. DOI: 10.3724/SP.J.1001.2008.00640.

[7] 陈玉明, 李伟. 粒向量与K近邻粒分类器[J]. 计算机研究与发展, 2019, **56**(12): 2600−2611. DOI: 10.7544/issn1000-1239.2019.20180572.
CHEN Y M, LI W. Granular Vectors and K Nearest Neighbor Granular Classifiers[J]. *J Comput Res Dev*, 2019, **56**(12): 2600−2611. DOI: 10.7544/issn1000-1239.2019.20180572.

[8] 张清华, 肖嘉瑜, 艾志华, 等. 自适应半径选择的近邻邻域分类器[J]. 模式识别与人工智能, 2022, **35**(11): 989−998. DOI: 10.16451/j.cnki.issn1003-6059.202211004.
ZHANG Q H, XIAO J Y, AI Z H, *et al*. Near Neighborhood Classifier with Adaptive Radius Selection[J]. *Pattern Recognit Artif Intell*, 2022, **35**(11): 989−998. DOI: 10.16451/j.cnki.issn1003-6059.202211004.

[9] 钞娜, 万仁霞, 苗夺谦. 直觉模糊信息系统下基于优势关系的邻域粗糙集[J]. 山西大学学报(自然科学版), 2023, **46**(1): 62−68. DOI: 10.13451/j.sxu.ns.2022086.
CHAO N, WAN R X, MIAO D Q. Neighborhood Rough Set Based on Dominant Relation in Intuitionistic Fuzzy Information System[J]. *J Shanxi Univ Nat Sci Ed*, 2023, **46**(1): 62−68. DOI: 10.13451/j.sxu.ns.2022086.

[10] 曾艺祥, 林耀进, 范凯钧, 等. 基于层次类别邻域粗糙集的在线流特征选择算法[J]. 南京大学学报(自然科学), 2022, **58**(3): 506−518.
ZENG Y X, LIN Y J, FAN K J, *et al*. Online Streaming Feature Selection Method Based on Hierarchical Class Neighborhood Rough Set[J]. *J Nanjing Univ Nat Sci*, 2022, **58**(3): 506−518.

[11] 杨洁, 王国胤, 李帅. 基于边界域的邻域知识距离度量

模型[J]. 计算机科学, 2020, **47**(3): 61−66. DOI: 10.11896/jsjkx.190500174.
YANG J, WANG G Y, LI S. Neighborhood Knowledge Distance Measure Model Based on Boundary Regions [J]. *Comput Sci*, 2020, **47**(3): 61−66. DOI: 10.11896/jsjkx.190500174.

[12] 陈玉明, 蔡国强, 卢俊文, 等. 一种邻域粒K均值聚类方法[J]. 控制与决策, 2023, **38**(3): 857−864. DOI: 10.13195/j.kzyjc.2021.1553.
CHEN Y M, CAI G Q, LU J W, *et al*. A Neighborhood Granular K-means Clustering Method[J]. *Control Decis*, 2023, **38**(3): 857−864. DOI: 10.13195/j.kzyjc.2021.1553.

[13] 王忠, 折延宏, 郑逸. 基于层次标签数据的模糊决策树构造算法[J]. 郑州大学学报(理学版), 2022, **54**(2): 24−31. DOI: 10.13705/j.issn.1671−6841.2021199.
WANG Z, ZHE Y H ,ZHENG Y. Fuzzy Decision Tree Construction Algorithm Based on Data with Hierarchical Labels[J]. *J Zhengzhou Univ Nat Sci Ed*, 2022, **54**(2): 24−31. DOI: 10.13705/j.issn.1671−6841.2021199.

[14] 王雅辉, 钱宇华, 刘郭庆. 基于模糊优势互补互信息的有序决策树算法[J]. 计算机应用, 2021, **41**(10): 2785−2792. DOI: 10.11772/j.issn.1001-9081.2020122006.
WANG Y H, QIAN Y H, LIU G Q. Ordinal Decision Tree Algorithm Based on Fuzzy Advantage Complementary Mutual Information[J]. *J Comput Appl*, 2021, **41**(10): 2785−2792. DOI: 10.11772/j. issn. 1001-9081.2020122006.

[15] ABRAMOVICH F, GRINSHTEIN V, LEVY T. Multiclass Classification by Sparse Multinomial Logistic Regression[J]. *IEEE Trans Inf Theory*, 2021, **67**(7): 4637−4646. DOI: 10.1109/TIT.2021.3075137.

[16] SHA M. A Cloud Based Sentiment Analysis Through Logistic Regression in AWS Platform[J]. *Comput Syst Sci Eng*, 2023, **45**(1): 857−868. DOI: 10.32604/csse.2023.031321.

[17] CASTILLA E, CHOCANO P J. A New Robust Approach for Multinomial Logistic Regression with Complex Design Model[J]. *IEEE Trans Inf Theory*, 2022, **68**(11): 7379−7395. DOI: 10.1109/TIT.2022.3187063.

[18] 王敬, 张宝, 谢晓, 等. 基于广义稀疏逻辑回归的全脑分类[J]. 信阳师范学院学报(自然科学版), 2022, **35**(3): 488−493. DOI: 10.3969/j.issn.1003-0972.2022.03.024.
WANG J, ZHANG B, XIE X, *et al*. Whole-brain Classification Based on Generalized Sparse Logistic Regression[J]. *J Xinyang Norm Univ Nat Sci Ed*, 2022, **35**(3): 488−493. DOI: 10.3969/j.issn.1003-0972.2022.03.024.

[19] PAN X L, XU Y T. A Safe Feature Elimination Rule for L_1-regularized Logistic Regression[J]. *IEEE Trans Pattern Anal Mach Intell*, 2022, **44**(9): 4544−4554. DOI: 10.1109/TPAMI.2021.3071138.

[20] LIU D, LI T R, LIANG D C. Incorporating Logistic Regression to Decision-theoretic Rough Sets for Classifications[J]. *Int J Approx Reason*, 2014, **55**(1): 197−210. DOI: 10.1016/j.ijar.2013.02.013.

[21] DIAO Y N, CHEN Q Q, LIU Y, *et al*. A Fuzzy Granular Logistic Regression Algorithm for SEMG-based Cross-individual Prosthetic Hand Gesture Classification[J]. *J Neural Eng*, 2023, **20**(2): 026029. DOI: 10.1088/1741-2552/acc42a.

[22] DUA D, GRAFF C. UCI Machine Learning Repository [EB/OL]. University of California, 2019. http://archive.ics.uci.edu/ml.