·人工智能与模式识别·　　文章编号：1000-3428(2022)07-0122-08　　文献标志码：A　　中图分类号：TP301

基于隶属度的模糊加权k近质心近邻算法

刘　利，张德生，肖燕婷

（西安理工大学 理学院，西安 710054）

摘　要：模糊k近质心近邻算法（FKNCN）的分类结果易受噪声点和离群点影响，并且算法对所有样本特征同等对待，不能体现样本特征的差异性。针对这两个问题，提出基于隶属度的模糊加权k近质心近邻算法MRFKNCN。利用密度聚类思想构造新的隶属度函数计算训练样本的隶属度，以减小噪声或离群样本对分类结果的影响。在此基础上，设计基于冗余分析的Relief-F算法计算每个特征的权重，删去较小权重所对应的特征和冗余特征，并通过加权欧氏距离选取有代表性的k个近质心近邻，提高分类性能。最终，根据最大隶属度原则确定待分类样本的类别。利用UCI和KEEL中的多个数据集对MRFKNCN算法进行测试，并与KNN、KNCN、LMKNCN、FKNN、FKNCN2和BMFKNCN算法进行比较。实验结果表明，MRFKNCN算法的分类性能明显优于其他6个对比算法，平均准确率最高可提升4.68个百分点。

关键词：k近质心近邻算法；隶属度；冗余分析；特征选择；数据分类

开放科学（资源服务）标志码（OSID）：

中文引用格式：刘利，张德生，肖燕婷.基于隶属度的模糊加权k近质心近邻算法[J].计算机工程，2022，48(7)：122-129.

英文引用格式：LIU L，ZHANG D S，XIAO Y T.Fuzzy weighted k-nearest centroid neighbor algorithm based on membership[J].Computer Engineering，2022，48(7)：122-129.

Fuzzy Weighted k-Nearest Centroid Neighbor Algorithm Based on Membership

LIU Li，ZHANG Desheng，XIAO Yanting

（School of Sciences，Xi'an University of Technology，Xi'an 710054，China）

【Abstract】The classification results of Fuzzy K-Nearest Centroid Neighbor（FKNCN）algorithm is susceptible to noise points，outliers，at the same time，the algorithm treats all sample features equally and cannot reflect the difference of sample features. To solve these two problems，fuzzy weighted k-nearest centroid neighbor algorithm（MRFKNCN）based on membership was proposed. Firstly，a new membership function is constructed by the idea of density clustering and the membership degree of training samples is calculated，which can avoid the influence of noise or outlier samples on the classification results. Then，the weight of each feature was calculated by the Relief-F algorithm of redundancy analysis，the features and redundant features corresponding to smaller weights were deleted，and k representative nearest centroid neighbors were selected by weighted Euclidean distance to improve the performance of classification. Finally，the classification of samples to be classified is determined by the maximum membership principle. The MRFKNCN algorithm is tested using multiple datasets in UCI and KEEL，and compared with KNN，KNCN，LMKNCN，FKNN，FKNCN2 and BMFKNCN. The experimental results show that the classification performance of MRFKNCN algorithm is significantly better than the other six comparison algorithms，the average accuracy can be improved by up to 4.68 percentage points.

【Key words】Fuzzy k-Nearest Centroid Neighbor（FKNCN）algorithm；membership；redundancy analysis；feature selection；data classification

DOI：10. 19678/j. issn. 1000-3428. 0062092

0　概述

分类任务[1]是利用已知类别的样本通过建立模型预测新样本的类别[2]。分类是数据挖掘中最重要的任务之一，应用十分广泛。目前，常用的分类算法主要有决策树[3]、贝叶斯分类器[4]、人工神经网络[5]、支持向量机[6]、k近邻（k-Nearest Neighbor，KNN）算法[7]等。这些已有的分类算法及其改进算法被广泛地应用于各个领域，如统计学[8]、医学[9]、模式识别[10]、决策理论[11]等。KNN算法的基本思想[12]是通

基金项目：国家自然科学基金青年科学基金项目（11801438）。

作者简介：刘　利（1997—），女，硕士研究生，主研方向为数据挖掘、分类分析；张德生，教授、博士；肖燕婷，副教授、博士。

收稿日期：2021-07-14　　**修回日期：**2021-08-27　**E-mail：**1257190813@qq.com

过已知类别的训练样本寻找待测样本的k个近邻，将k个近邻中出现频率最高的类别作为待测样本的类别。由于KNN算法具有理论简单、易于操作等优点，被认为是数据挖掘中最简单的方法之一[13]。但是，KNN算法也存在不足：第1个问题是它没有考虑样本的分布，当样本分布不均匀或样本中存在噪声样本时，分类精度会明显下降；第2个问题是所有训练样本具有同等重要性，判断待分类样本的类别时没有考虑k个近邻的区别；第3个问题是使用单一的多数投票原则进行分类决策。以上3个问题都会影响分类的准确率及效率。

针对KNN算法存在的问题，很多研究者提出了不同的改进算法。文献[14]提出了k近质心近邻(k-Nearest Centroid Neighbor，KNCN)算法，该算法的设计思想是近邻点要尽可能地离待分类样本近，而且近邻点要均匀地分布在待分类样本周围，通过这些近邻点所属类别判断待测样本的类别。文献[15]提出了一种基于局部权重的k近质心近邻算法LMKNCN，该算法的设计思想是为k个近质心近邻赋予不同的权重，再通过决策函数对每个测试样本进行分类。文献[16]将模糊理论与KNN相结合，提出了模糊k近邻(Fuzzy k-Nearest Neighbor，FKNN)算法，该算法的基本思想是为训练样本分配隶属度，根据k个近邻样本的隶属度和距离的权重，通过最大隶属度原则确定待分类样本的类别。文献[17]提出了自适应k值的FKNN算法，该算法的设计思想是通过改进的粒子群算法自适应优化k值和模糊强度参数m，从而提高算法性能。文献[18]将KNCN算法和FKNN算法相结合，提出了模糊k近质心近邻(Fuzzy KNCN，FKNCN)算法，该算法的设计思想是使用近质心近邻的概念来确定最近邻，并使用模糊理论为每个类分配隶属度，同时解决了样本分布和权重问题。文献[19]提出了基于Bonferroni均值的FKNCN算法，该算法的设计思想是运用均值算子和近质心近邻概念，计算最近的局部Bonferroni均值向量确定待分类样本的类别标签，该算法对异常值具有很好的鲁棒性。

然而，上述FKNCN算法及其改进算法在计算训练样本的隶属度时没有考虑训练样本中存在的噪声点或离群点，这在小样本数据集中会大幅降低分类精度。同时，算法忽略了训练样本特征的差异性，没有进行特征选择，影响了分类性能。针对这2个问题，本文对FKNCN算法进行改进，提出基于隶属度的模糊加权k近质心近邻算法MRFKNCN。设计新的隶属度函数计算训练样本的隶属度，区分训练样本中存在的噪声或离群样本与有效样本。在此基础上，通过基于冗余分析的Relief-F算法计算每个特征的权重，删去较小权重所对应的特征和冗余特征，选出重要特征后根据加权欧氏距离选取k个有代表性的近质心近邻，并确定待测样本的类标号。

1　预备知识

1.1　符号说明

本文算法所使用的符号说明见表1。

表1　符号说明

Table 1　Symbol description

符号	说明
$T=\{x_j\in\mathbb{R}^p\}_{j=1}^n$	训练样本集
$\{y_i\mid y_1,y_2,\cdots,y_m\}$	待测样本集
$X^{\mathrm{NCN}}=\{x_r^{\mathrm{NCN}}\in\mathbb{R}^P\}_{r=1}^k$	k个近质心近邻集合
$x_{jl}=\{x_{j1},x_{j2},\cdots,x_{jp}\}$	第j个训练样本的第l个特征
$C=\{c_1,c_2,\cdots,c_M\}$	类标签集
n	训练样本个数
M	类别个数
P	特征个数
k	近邻个数

1.2　FKNCN算法

假定在p维特征空间中，训练样本集$T=\{x_j\in\mathbb{R}^p\}_{j=1}^n$有$M$个类标签$c_1,c_2,\cdots,c_M$，给定一个待测样本点$y$，FKNCN算法步骤如下：

步骤1　利用式(1)计算待测样本y与所有训练样本间的欧氏距离，进行升序排列，选择最短距离所对应的训练样本作为第1个近质心近邻点x_1^{NCN}：

$$d(y,x_j)=\sqrt{(y-x_j)^{\mathrm{T}}(y-x_j)} \tag{1}$$

步骤2　当$r=2,3,\cdots,k$时，利用式(2)计算T_i(T_i表示除去被选为近质心近邻的训练样本之外所有剩余的训练样本)中的训练样本与之前所选的$r-1$个近质心近邻的质心：

$$x_{rj}^{\mathrm{c}}=\frac{(x_1^{\mathrm{NCN}}+x_2^{\mathrm{NCN}}+\cdots+x_{r-1}^{\mathrm{NCN}})+x_j}{r} \tag{2}$$

步骤3　利用式(3)计算训练样本的质心与待测样本间的最短距离，选取所对应的训练样本作为第r个近质心近邻：

$$d(y,x_{rj}^{\mathrm{c}})=\min\sqrt{(y-x_{rj}^{\mathrm{c}})^{\mathrm{T}}(y-x_{rj}^{\mathrm{c}})} \tag{3}$$

步骤4　记k个近质心近邻的集合为$X^{\mathrm{NCN}}=\{x_r^{\mathrm{NCN}}\in\mathbb{R}^p\}_{r=1}^k$，利用式(4)计算待测样本$y$属于第$i$类的模糊隶属度：

$$u_i^{\mathrm{NCN}}(y)=\frac{\sum_{r=1}^{k}u_{ir}(1/d(y,x_r^{\mathrm{NCN}})^{2/(m-1)})}{\sum_{r=1}^{k}(1/d(y,x_r^{\mathrm{NCN}})^{2/(m-1)})} \tag{4}$$

其中：$i=c_1,c_2,\cdots,c_M$，$d(y,x_r^{\mathrm{NCN}})$为待测样本和近质心近邻的欧氏距离；m为模糊强度参数，代表每个近质心近邻与待测样本的权重；u_{ir}为训练样本的隶属度。有以下2种计算方式定义u_{ir}[18]：一种是硬性分类，隶属度定义如式(5)所示；另一种是使用模糊隶属度，找到每个近质心近邻训练样本的k个近邻(x_{rk}^{NCN})，则x_{rk}^{NCN}属于第i类的隶属度定义如式(6)所示。其中：n_r表示第r个近质心近邻训练样本的k个近邻属于第i类的近邻个数；$c(x_{rk}^{\mathrm{NCN}})$表示近质心近邻训练样本的k个近邻所属的类别。

$$u_{ir}(x_r^{\mathrm{NCN}})=\begin{cases}1, & x_r^{\mathrm{NCN}}\in i\\ 0, & x_r^{\mathrm{NCN}}\notin i\end{cases} \tag{5}$$

$$u_{ir}(x_{rk}^{\mathrm{NCN}})=\begin{cases}0.51+0.49(n_r/k), & c(x_{rk}^{\mathrm{NCN}})=i\\ 0.49(n_r/k), & c(x_{rk}^{\mathrm{NCN}})\neq i\end{cases} \tag{6}$$

利用式(5)和式(6)均可计算训练样本的隶属度。一般而言，优先选择式(6)计算其隶属度[18]，原因在于该计算方式引入了模糊理论，将训练样本的隶属度模糊化，通过训练样本的k个近邻确定其隶属度，能够确保训练样本在自己的类中被赋予较高的权重，而在其他类中被赋予较低的权重。

步骤5 通过最大的模糊隶属度值判断待测样本y的所属类别，如式(7)所示：

$$y=\arg\max(u_i^{\mathrm{NCN}}(y)) \tag{7}$$

步骤6 对于一个新的待测样本点，重复步骤1~步骤5。

1.3 特征的相关性度量

定义1 皮尔逊相关系数[20]

设p维空间中的2个样本点$e=(e_1,e_2,\cdots,e_p)$，$f=(f_1,f_2,\cdots,f_p)$，两者之间的皮尔逊相关系数的计算公式如式(8)所示：

$$r_{ef}=\frac{\sum_{i=1}^{p}(e_i-\bar{e})(f_i-\bar{f})}{\sqrt{\sum_{i=1}^{p}(e_i-\bar{e})}\sqrt{\sum_{i=1}^{p}(f_i-\bar{f})}} \tag{8}$$

其中：$\bar{e}=\frac{1}{p}\sum_{i=1}^{p}e_i$，$\bar{f}=\frac{1}{p}\sum_{i=1}^{p}f_i$，分别表示样本$e$、$f$所有特征的均值。

r_{ef}表示特征之间的线性相关性，其取值范围是$[-1,1]$，即$0\leqslant|r_{ef}|\leqslant 1$，相关系数的绝对值越大，相关性越强，相关系数的绝对值越接近于0，相关度越弱。

2 MRFKNCN算法

隶属度函数的设计是FKNCN算法的关键，但FKNCN算法在计算训练样本的隶属度时，并没有考虑噪声点或离群点对分类的影响，同时该算法在计算待测样本与训练样本间的欧氏距离时，把所有训练样本的各维特征等同对待，没有区分它们的重要程度，这些都会影响分类的性能。为此，本文提出基于隶属度的模糊加权k近质心近邻算法MRFKNCN。通过密度聚类思想为训练样本设计新的隶属度函数、利用基于冗余分析的Relief-F算法计算特征的权重、确定待测样本的类别这3个部分克服噪声点、离群点的影响，同时解决相同特征权重的问题，提高分类的效率和准确率。

2.1 隶属度函数构造

样本集中经常会出现噪声点或离群点，而FKNCN算法在计算训练样本的隶属度时，没有区分这些样本与有效样本，导致所有训练样本的隶属度相同，这会在很大程度上影响分类的准确率。本文采用密度聚类的思想构造最小包围球。首先计算最小包围球的类中心和半径；然后根据训练样本在最小包围球的位置确定其隶属度；最后根据隶属度的大小判断训练样本是离群或噪声样本还是有效样本。计算最小包围球类中心和半径的具体步骤如下：

步骤1 计算训练样本x_j与训练样本集中其他样本的欧氏距离，找到x_j的第r个近邻$(x_{jr},r=1,2,\cdots,k)$。

步骤2 根据样本x_j的k个近邻构造$n\times k$的密度矩阵$\boldsymbol{D}=(d(x_j,x_{jr}))_{n\times k}$。

步骤3 利用式(9)计算所有样本的密度$\rho(x_j)$：

$$\rho(x_j)=\frac{z_j}{\sum_{j=1}^{n}z_j} \tag{9}$$

其中：$z_j=\sum_{r=1}^{k}\frac{1}{d(x_j,x_{jr})}$，为样本$x_j$与其$k$个近邻距离的倒数和。将密度从大到小排列，得到密度点集G。

步骤4 选择最大密度样本点$x_{\max}$，并找出离该点最近的样本点$x_{n_\max}$，通过这2个点确定最小包围球的类中心$O(T)$：

$$O(T)=0.7x_{\max}+0.3x_{n\_\max} \tag{10}$$

步骤5 利用式(10)计算最小包围球的半径：

$$R(T)=\lambda\frac{a(T)}{n^{\delta}} \tag{11}$$

其中：λ为自定义的惩罚因子；δ为半径调整系数；$a(T)$为类中心到所有样本点距离的平均值。

计算出样本集中最小包围球的类中心和半径后，利用式(12)确定训练样本的隶属度：

$$u(x_j)=\begin{cases}0.7\times\left(\dfrac{1-d(x_j)/R(T)}{1+d(x_j)/R(T)}\right)+0.3, & d(x_j)\leqslant R(T)\\ 0.3\times\left(\dfrac{1}{1+(d(x_j)-R(T))}\right), & d(x_j)>R(T)\end{cases} \tag{12}$$

其中：$d(x_j)$为训练样本x_j与最小包围球类中心$O(T)$的欧氏距离。从式(12)可以看出，训练样本离最小包围球的类中心越远，该训练样本的隶属度就越小。如果训练样本位于最小包围球之内，其隶属度都大于0.3；反之，位于最小包围球之外的训练样本，其隶属度都小于等于0.3。位于最小包围球之外的样本一般都为离群或噪声样本。本文构造最小包围球方法简单有效，原因在于其只需计算最小包围球的类中心和半径，再通过式(12)为离群或噪声样本赋予较小的隶属度，即可快速区分出离群或噪声样本与有效样本。

2.2　基于冗余分析的Relief-F算法

本节提出基于冗余分析的Relief-F特征选择算法计算每个特征的权重。首先将FKNCN算法中的欧氏距离改为特征加权的欧氏距离；然后通过加权欧氏距离确定待测样本的近质心近邻；最后通过近质心近邻的隶属度和距离加权确定待测样本的类隶属度。

2.2.1　特征权重的计算

在分类过程中，并不是所有的特征都与分类强相关，也会存在一些不相关特征及冗余特征。如果在分类时不处理这些特征，会出现计算成本高、分类性能低等问题。为了得到最优特征子集，特征选择是必不可少的。Relief-F算法[21]是最成功的特征选择方法之一，算法的具体步骤如下：

1)对所有特征归一化处理：

$$x'_{j,l}=\frac{x_{j,l}-\mu_l}{s_l} \tag{13}$$

其中：$x'_{j,l}$为样本x_j通过归一化处理后在第l个特征上的值；μ_l、s_l分别为所有样本在第l个特征上的均值、标准差；$x_{j,l}$为样本x_j在第l个特征上的值。

2)从训练集中随机选择样本点x_j，找出x_j同类的k个最近邻样本集和不同类的k个样本集。

3)通过式(14)计算每个特征的特征权重[22]：

$$w(l)=w(l)-\sum_{r=1}^{k}\mathrm{diff}(A_l,x_j,H_r)/\alpha k+\sum_{c\neq\mathrm{class}(x_j)}\left[\frac{p(c)}{1-p(\mathrm{class}(x_j))}\sum_{r=1}^{k}\mathrm{diff}(A_l,x_j,M_r(c))\right]/\alpha k \tag{14}$$

其中：$\mathrm{diff}(A_l,x_j,H_r)$表示样本$x_j$与$H_r$在第$l$个特征上的距离；$\mathrm{diff}(A_l,x_j,M_r(c))$表示样本$x_j$与$M_r(c)$在第$l$个特征上的距离；$p(c)$表示属于类别$c$的样本出现的概率。

虽然Relief-F算法在处理多分类问题时效率高并且能够很好地剔除不相关特征，但不能过滤冗余特征。为此，本文在Relief-F算法的基础上提出基于冗余分析的Relief-F特征选择算法计算所有特征的权重，算法描述如下：

算法1　基于冗余分析的Relief-F算法

输入　训练集T，样本抽样次数α，最近邻样本个数k

输出　每个特征的特征权重w

步骤1　所有特征归一化处理。

步骤2　将所有特征权重置0。

步骤3　在T中随机选择样本点x_j。

步骤4　找到与x_j同类的k个最近邻样本集H_r。

步骤5　每个类$c\neq\mathrm{class}(x_j)$，找到与x_j不同类的k个最近邻样本集$M_r(c)$。

步骤6　更新每个特征的特征权重。

步骤7　根据特征权重阈值，选择分类权重最大的特征集合。

步骤8　冗余分析。利用皮尔森相关系数计算特征之间的相关性。

以上步骤是在一次抽样下计算每个特征的特征权重，经过α次抽样后，将特征权重更新α次，并设置一个特征权重阈值Γ，将每个特征权重与总特征权重的比值累积，选择累积特征权重比大于阈值Γ的特征作为新特征，并删掉剩余的不相关特征，然后分析新特征之间的相关性，在特征相关性较大的情况下保留权重较高的特征，消除权重较低的特征，目的是消除冗余特征的干扰。通过基于冗余分析的Relief-F算法计算特征权重，同时减少特征的数量、降低维数，从而完成特征选择。

2.2.2　加权欧氏距离的计算

利用式(15)计算待测样本y与训练样本x_j的加权欧氏距离，从而确定待测样本的k个近质心近邻：

$$d(y,x_j)=\sqrt{\sum_{l=1}^{p}w(l)(y_l-x_{jl})^{\mathrm{T}}(y_l-x_{jl})} \tag{15}$$

其中：$w(l)$为训练样本第l个特征的特征权重，且$w(l)>0$，$\sum_{l=1}^{p}w(l)=1$。

2.2.3　待分类样本类别的确定

计算出每个特征的特征权重后，利用式(16)计算待测样本y属于每个类别的隶属度值：

$$u_i^{\mathrm{NCN}}(y)=\frac{\sum_{r=1}^{k}u_i(x_r^{\mathrm{NCN}})(1/d(y_l,x_{rl}^{\mathrm{NCN}})^{2/(m-1)})}{\sum_{r=1}^{k}(1/d(y_l,x_{rl}^{\mathrm{NCN}})^{2/(m-1)})} \tag{16}$$

其中：$d(y_l,x_{rl}^{\mathrm{NCN}})$表示待测样本$y$和第$r$个近质心近邻的加权欧氏距离；$u_i(x_r^{\mathrm{NCN}})$表示待测样本的第$r$个近质心近邻属于第$i$类的隶属度。若$x_r^{\mathrm{NCN}}$属于类别$i$，则利用式(12)计算$u_i(x_r^{\mathrm{NCN}})$；若不属于类别$i$，则$u_i(x_r^{\mathrm{NCN}})$的值为0。

得到待测样本y属于每个类别的隶属度后，通过最大隶属度原则确定待测样本y的类别。

2.3　算法描述

MRFKNCN算法的设计思想为：首先计算训练样本的隶属度；然后计算所有特征的权重，并找出

k个近质心近邻;最后计算待测样本的模糊隶属度,通过最大隶属度原则确定待测样本的类别。具体步骤如算法2所示,算法流程如图1所示。

算法2 MRFKNCN算法

输入 近质心近邻个数k,待测样本点y,训练样本集T,模糊强度参数m

输出 待测试样本y的类别

步骤1 利用式(9)计算训练样本的密度。

步骤2 利用式(10)和式(11)计算最小包围球的类中心和半径。

步骤3 利用式(12)计算训练样本隶属度。

步骤4 通过基于冗余分析的Relief-F算法计算所有特征的权重。

步骤5 利用式(15)计算待测样本与训练样本之间的加权欧氏距离,找出k个近质心近邻集合。

步骤6 利用式(16)计算待测样本的模糊隶属度。

步骤7 根据最大隶属度原则确定待分类样本的类别。

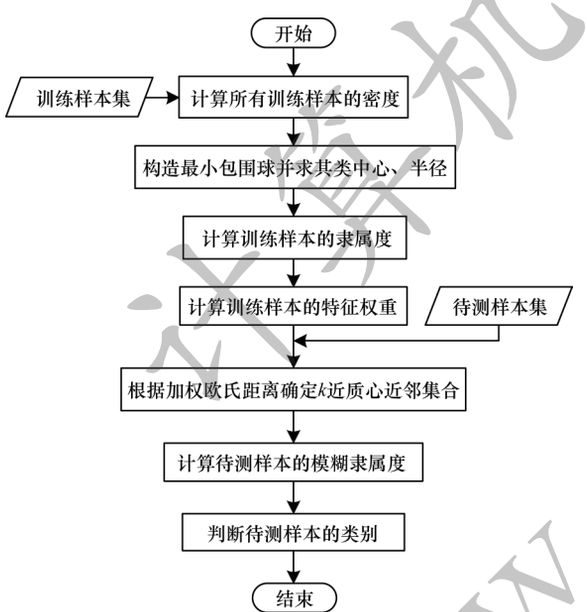

图1 MRFKNCN算法流程

Fig.1 Procedure of MRFKNCN algorithm

2.4 算法复杂度分析

假设n表示数据集的规模,p表示特征维数,k表示近邻个数,M表示类别个数。基于隶属度的模糊k近质心近邻算法的时间复杂度主要来源于以下5个部分:1)通过密度聚类思想计算训练样本的隶属度,时间复杂度为$O(np)$;2)计算每个特征的特征权重,时间复杂度为$O(np\mu)$;3)计算训练样本的质心,时间复杂度为$O(nk)$;4)待测样本到各个类的加权距离,时间复杂度为$O(np)$;5)通过最大隶属度原则确定待测样本的类别,时间复杂度为$O(M)$。因此,MRFKNCN算法总的时间复杂度为$O(2np+nk)$。

3 仿真实验与分析

3.1 数据集

为验证本文MRFKNCN算法的有效性,选用UCI和KEEL中的11个标准数据集和4个含噪数据集进行仿真实验,所有实验都在Matlab2014b的环境下完成。表2、表3列出了实验中所用数据集的相关信息。

表2 标准数据集

Table 2 Standard data sets

数据集	样本数	类别数	维数	测试样本数
Zoo	101	7	18	20
Hayes-roth	160	3	4	32
Ecoli	336	8	8	67
Glass	214	9	7	43
Sends	210	7	3	42
Thyroid	7 200	3	21	720
Balance	625	3	4	63
Segment	2 310	7	18	231
Movement	360	15	90	72
Arrhythmia	452	3	279	74
Multivariate	7 797	5	617	780

表3 含噪数据集

Table 3 Datas sets with noise

数据集	样本数	类别数	维数	噪声样本数	测试样本数
Iris	150	3	4	3	30
Vehicle	846	4	18	16	85
Wine	178	3	13	3	37
Letter	7 648	10	16	152	765

3.2 算法参数设置

为了更好地测试各算法的分类效果,对本文算法所使用的相关参数进行调优。

参考文献[23],取$\lambda=2$,$\delta=0.14$计算最小包围球的半径。计算完特征权重后,需要设置一个阈值$\Gamma\in(0,1)$[22]。对本文的数据集进行多次试验可知,当$\Gamma=0.8$时,选出的新特征最具有代表性,模糊强度参数$m=2$,当模糊隶属度值与距离平方成反比时,在分类过程中会得到最优结果[18]。

3.3 实验结果与分析

本节设计了4个实验来验证MRFKNCN算法的有效性,将分类的准确率作为评价标准,比较本文算法与其他算法的性能。准确率的计算方法如下:

$$A=\frac{N_c}{N_t}\times 100\% \tag{17}$$

其中:N_c为正确分类的样本个数;N_t为实际分类的样本个数。

对于样本总数较小的数据集,通过10次5折交叉验证进行实验;对于样本总数大的数据集,通过10次10折交叉验证进行实验,最后将所有实验得到的准确

率平均值作为测试结果。实验1~实验3均采用交叉验证法确定最优k值。

3.3.1　MRFKNCN算法总体性能分析

实验1　为验证本文所提的新的隶属度函数在噪声点或离群点影响下的有效性，将MRFKNCN算法与利用式(5)计算隶属度的FKNCN算法(命名为FKNCN1)及利用式(6)计算隶属度的FKNCN算法(命名为FKNCN2)进行比较，运用表3含噪数据集。实验结果如表4所示。

表4　MRFKNCN与FKNCN算法的平均准确率

Table 4　Average accuracy of MRFKNCN and FKNCN algorithms　%

数据集	FKNCN1算法	FKNCN2算法	MRFKNCN算法
Iris	92.35	93.57	96.83
Vehicle	71.78	71.02	74.94
Wine	83.54	84.49	87.18
Letter	93.35	93.88	96.21

表4结果表明，当训练集中含有噪声点或离群点时，MRFKNCN算法的平均准确率明显高于FKNCN1算法和FKNCN2算法，在Iris、Vehicle、Wine、Letter这4个含噪数据集上，MRFKNCN算法的平均准确率比FKNCN1算法分别提高4.48、3.16、3.64、2.86个百分点，比FKNCN2算法分别提高3.26、3.92、2.69、2.33个百分点，这表明本文所设计的新隶属度函数可以很好地识别出训练样本集中的噪声点或离群点，尤其是在Iris小数据集中，MRFKNCN算法获得了较高的准确率。

实验2　为验证基于冗余分析的Relief-F算法计算特征权重方法的有效性，将MRFKNCN算法与未加权欧氏距离的MRFKNCN算法(命名为MRFKNCN_N)、确定统一特征权重的MRFKNCN算法(命名为MRFKNCN_U)进行对比，运用表3中Arrhythmia、Segment、Zoo、Balance、Thyroid这5个数据集，3种算法的平均准确率结果如图2所示。

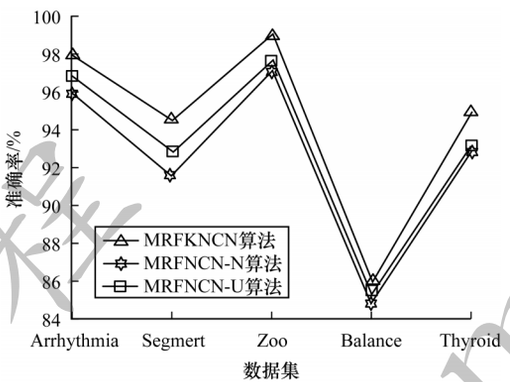

图2　3种算法平均准确率对比

Fig.2　Comparison of average accuracy of three algorithms

图2结果表明，5个数据集中MRFKNCN的分类准确率都明显优于MRFKNCN_N算法和MRFKNCN_U算法，说明不同的特征有不同的贡献率。因此，为了保证算法的准确率，应分别确定每个样本特征的权重，区分其差异，从而提高分类的性能。

实验3　在最优k值下比较MRFKNCN、KNN[13]、KNCN[14]、LWKNCN[15]、FKNN[16]、FKNCN[18]和BMFKNCN[19]算法的分类平均准确率，所得结果如表5所示，其中加粗数字为最优值。

表5　最优k值下MRFKNCN与其他6种算法的平均准确率

Table 5　Average accuracy of MRFKNCN and other six algorithms under optimal k value　%

数据集	KNN算法	KNCN算法	LWKNCN算法	FKNN算法	FKNCN2算法	BMFKNCN算法	MRFKNCN算法
Zoo	92.73	94.58	95.46	95.62	97.25	98.57	**98.93**
Hayes-roth	70.35	71.29	73.87	72.93	73.44	74.21	**74.52**
Ecoli	81.18	82.36	82.93	82.74	83.32	84.29	**84.48**
Glass	69.85	72.44	74.57	73.13	75.96	74.52	**76.35**
Sends	88.07	88.36	89.99	88.81	89.25	90.45	**92.41**
Thyroid	92.54	93.38	93.89	93.23	94.02	**94.97**	94.03
Balance	81.27	82.02	84.23	83.58	84.35	85.78	**86.01**
Segment	89.07	89.92	92.85	90.67	91.53	93.33	**94.48**
Movement	80.26	80.38	82.52	81.81	82.45	83.06	**85.98**
Arrhythmia	94.26	95.33	96.01	95.54	96.37	96.42	**98.05**
Multivariate	79.51	79.67	80.46	78.95	80.14	81.79	**85.28**
平均准确率	83.55	84.52	86.07	85.18	86.19	87.03	**88.23**

表5结果表明，虽然在Thyroid数据集中，BMFKNCN算法取得了较高的准确率，但是MRFKNCN算法的准确率仍高于其他5种对比算法。MRFKNCN算法在其余10个数据集中的准确率都高于其他6种对比算法的准确率，尤其是在Movement、Arrhythmia、Multivariate这3个高维数据集上的准确率大幅提升，说明MRFKNCN算法不仅可以去除噪声样本对分类性能的影响，还可以选出有代表性的特征提高分类的准确率。同时，MRFKNCN算法在11个数据集中获得了最高的平均准确率。

3.3.2 MRFKNCN算法在不同k值下的性能分析

实验4 为了验证MRFKNCN算法在不同k值下的分类性能，将MRFKNCN算法与6个对比算法在k=1~15时进行比较。运用表3的Hayes-roth、Ecoli、Glass、Sends、Cleveland这5个数据集。图3~图7给出了7种算法的分类准确率对比折线图。

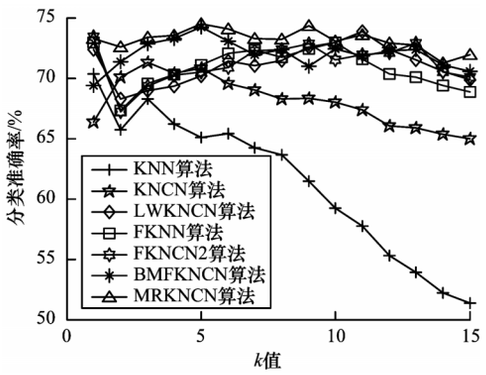

图3 在Hayes-roth数据集上的实验结果

Fig.3 Experimental results on the Hayes-roth dataset

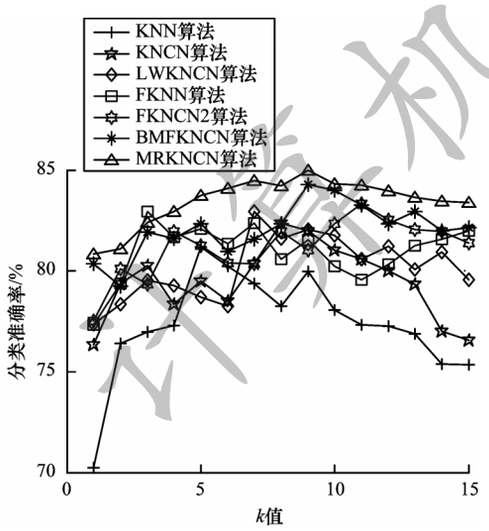

图4 在Ecoli数据集上的实验结果

Fig.4 Experimental results on the Ecoli dataset

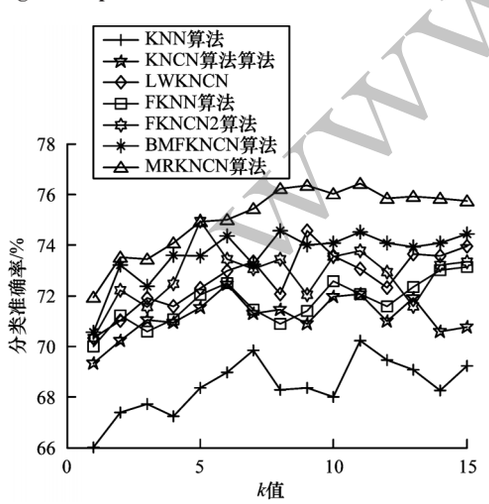

图5 在Glass数据集上的实验结果

Fig.5 Experimental results on the Glass dataset

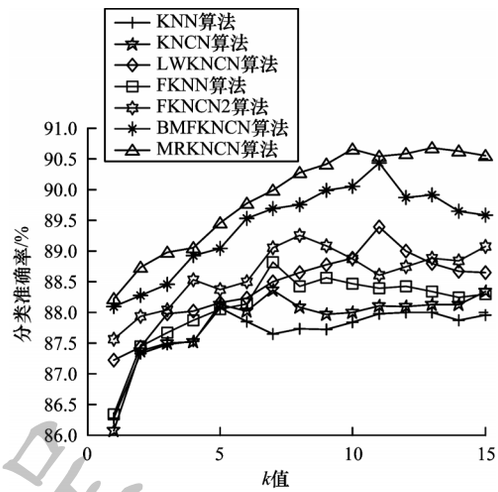

图6 在Sends数据集上的实验结果

Fig.6 Experimental results on the Sends dataset

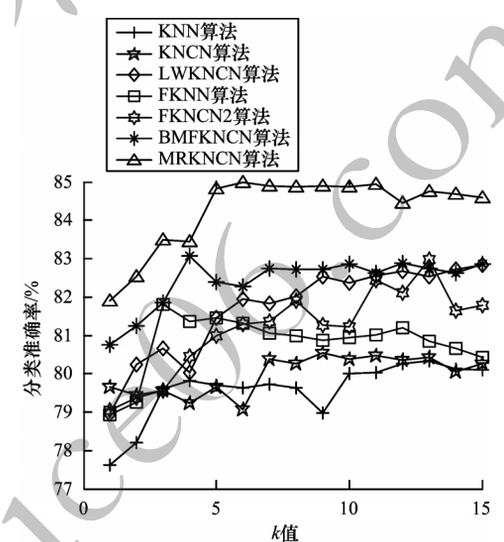

图7 在Movement数据集上的实验结果

Fig.7 Experimental results on the Movement dataset

4 结束语

针对FKNCN算法未区别样本特征的不足，本文提出基于隶属度的模糊加权k近质心近邻算法MRFKNCN。利用密度聚类思想设计新的隶属度函数计算训练样本的隶属度，通过基于冗余分析的Relief-F算法计算每个特征的权重，删去不相关特征和冗余特征，选出重要的特征，并利用加权欧氏距离选取k个近质心近邻，最终根据最大隶属度原则对待测样本进行分类。该算法有效解决了训练样本中存在噪声样本或离群样本的问题，而且还为每个特征赋予不同的权重，更符合分类的实际情况。实验结果表明，MRFKNCN算法在分类性能上明显优于其他对比算法。由于FKNCN算法对参数k和模糊强度因子m敏感也会影响分类性能，因此下一步将研究如何自适应地优化FKNCN算法的参数k和模糊强度因子m，进一步提高算法准确率。

参考文献

[1] SYAHARA Z, ADIHA R N, WINDARTO A P. Implementasi data mining algoritma apriori pada sistem persediaan bahan bangunan di karang sari[J]. BRAHMANA Jurnal Penerapan Kecerdasan Buatan, 2021, 2(2): 107-115.

[2] 王谟瀚，翟俊海，齐家兴. 基于MapReduce和Spark的大规模压缩模糊K-近邻算法[J]. 计算机工程，2020，46(11)：139-147.
WANG M H, ZHAI J H, QI J X. Large-scale condensed fuzzy K-nearest neighbor algorithm based on MapReduce and Spark[J]. Computer Engineering, 2020, 46(11): 139-147. (in Chinese)

[3] GUERRERO M D, VANDERLOO L M, RHODES R E, et al. Canadian children's and youth's adherence to the 24-h movement guidelines during the COVID-19 pandemic: a decision tree analysis[J]. Journal of Sport and Health Science, 2020, 9(4): 313-321.

[4] ABBAS A, SHIHAB M. Diagnosis the breast cancer using Bayesian rough set classifier[J]. Iraqi Journal of Science, 2017, 58(1B): 302-308.

[5] DONG Y L, WANG H M. Robust output feedback stabilization for uncertain discrete-time stochastic neural networks with time-varying delay[J]. Neural Processing Letters, 2020, 51(1): 83-103.

[6] HU J L, WANG J R, LIN J N, et al. MD-SVM: a novel SVM-based algorithm for the motif discovery of transcription factor binding sites[J]. BMC Bioinformatics, 2019, 20(Suppl 7): 200.

[7] MA Z F, TIAN H P, LIU Z C, et al. A new incomplete pattern belief classification method with multiple estimations based on KNN[J]. Applied Soft Computing, 2020, 90(4): 106-175.

[8] ANGKASA A, FITRIANAH D. The implementation of classification algorithm C4. 5 in determining the illness risk level for health insurance company in Indonesia[J]. International Journal of Computer Applications, 2020, 177(37): 44-50.

[9] BARWAL R K, RAHEJA N. Comparative analysis of classification methods based on medical datasets[J]. Journal of Computational and Theoretical Nanoscience, 2020, 17(6): 2737-2745.

[10] RAMACHANDRAN S K, MANIKANDAN P. An efficient ALO-based ensemble classification algorithm for medical big data processing[J]. International Journal of Medical Engineering and Informatics, 2021, 13(1): 54-58.

[11] ZHAN L Y, MA X H, FANG W Q, et al. A rapid classification method of aluminum alloy based on laser-induced breakdown spectroscopy and random forest algorithm[J]. Plasma Science and Technology, 2019, 21(3): 152-158.

[12] WANG F, YU Y L, WANG X K, et al. Residential electricity consumption level impact factor analysis based on wrapper feature selection and multinomial logistic regression[J]. Energies, 2018, 11(5): 1180-1193.

[13] DINESH KUMAR D S, RAO P V. Implementing and analysing FAR and FRR for face and voice recognition (multimodal) using KNN classifier[J]. International Journal of Intelligent Unmanned Systems, 2019, 8(1): 55-67.

[14] SÁNCHEZ J S, PLA F, FERRI F J. On the use of neighbourhood-based non-parametric classifiers[J]. Pattern Recognition Letters, 1997, 18(11/12/13): 1179-1186.

[15] 谢红，赵洪野，解武. 基于局部权重k-近质心近邻算法[J]. 应用科技，2015，42(5)：10-13.
XIE H, ZHAO H Y, XIE W. Local weight-based k-nearest centroid neighbor algorithm [J]. Applied Science and Technology, 2015, 42(5): 10-13. (in Chinese)

[16] KELLER J M, GRAY M R, GIVENS J A. A fuzzy K-nearest neighbor algorithm[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1985, 15(4): 580-585.

[17] JAMALUDDIN J, SIRINGORINGO R. Improved fuzzy K-nearest neighbor using modified particle swarm optimization [J]. Journal of Physics Conference Series, 2017, 930(1): 1-10.

[18] ROSDI B A, JAAFAR H, RAMLI D A. Finger vein identification using fuzzy-based k-nearest centroid neighbor classifier[C]//Proceedings of AIP Conference. Pahang, Malaysia: AIP Publishing, 2015: 649-654.

[19] WIDYADHANA A, PUTRA C B P, INDRASWARI R, et al. A bonferroni mean based fuzzy K nearest centroid neighbor classifier[J]. Jurnal Ilmu Komputer Dan Informasi, 2021, 14(1): 65-71.

[20] MENDHE P, BALPANDE R, KHOBRAGADE A. Fast fractal image encoding scheme based on Absolute value of Pearson correlation coefficient[C]//Proceedings of International Conference on Communications and Signal Processing. Washington D. C., USA: IEEE Press, 2015: 1036-1040.

[21] GAVISIDDAPPA G, MAHADEVAPPA S, PATIL C, et al. Multimodal biometric authentication system using modified ReliefF feature selection and multi support vector machine[J]. International Journal of Intelligent Engineering and Systems, 2020, 13(1): 1-12.

[22] SHAIK E K, KUMAR P R. Epilepsy identification based on VMD, RELIEFF algorithm and machine learning classification techniques[J]. International Journal of Recent Technology and Engineering, 2019, 8(3): 6180-6185.

[23] CHEN L C, GAO S, LIU B X, et al. FEW-NNN: a fuzzy entropy weighted natural nearest neighbor method for flow-based network traffic attack detection[J]. China Communications, 2020, 17(5): 151-167.

编辑　金胡考