DOI：10.7652/xjtuxb201302004

采用邻域决策分辨率的特征选择算法

诸文智，司刚全，张彦斌
（西安交通大学电气工程学院，710049，西安）

摘要：针对目前基于粗糙集模型的特征选择算法无法直接应用于数值型数据、必须经过离散化过程而造成决策信息丢失的问题，提出了一种基于邻域决策分辨率的特征选择算法。该算法根据邻域信息粒中决策分布与其分类能力间的关系，提出了邻域决策确定性（N_c）来衡量单个信息粒的决策分辨能力；并根据特征向量空间上所有信息粒所具有的 N_c 累加值，定义了邻域决策分辨率作为特征子集上决策可分辨性的量度，从而将名义型和数值型数据统一在同一特征选择算法框架下。仿真实验和实际应用的结果表明，该算法性能优于目前主流基于邻域粗糙集的特征选择方法。

关键词：特征选择；邻域粗糙集；邻域决策分辨率
中图分类号：TP274 **文献标志码**：A **文章编号**：0253-987X(2013)02-0020-08

Feature Selection Algorithm Based on Neighborhood Decision Distinguishing Rate

ZHU Wenzhi, SI Gangquan, ZHANG Yanbin
(School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: The current feature selection algorithms based on the neighborhood rough set (NRS) model are unable to evaluate numerical dataset directly, a discretization procedure becomes necessary to transform the datasets into discrete forms, but inevitably leads to useful decision information loss. To solve this difficulty, a feature selection algorithm based on the neighborhood effective information rate is proposed. In view point of granulated neighborhood, the relation between the decision discernibility and the decision distribution is analyzed, and the neighborhood decision certainty (N_c) is defined to indicate the degree of distinguishing capability in each individual neighborhood granule. The neighborhood decision distinguishing rate (NDDR) of the feature subset, which evaluates the ability of the subspace to approximate decision space, is established based on the sum of the N_c values of the information granules induced by the corresponding feature space. Then the nominal and numerical datasets can be integrated into the same feature selection algorithm framework. The simulation and application illustrate that the proposed algorithm outperforms the other NRS-based ones.

Keywords: feature selection; neighborhood rough set; neighborhood decision distinguishing rate

特征选择指在保持决策表分类能力不变的前提下，选择相关和必须的条件属性，删除不相关和冗余的条件属性。特征选择算法作为知识发现技术的重要环节之一，提高了后续知识学习算法的运算速度，增强了所得学习模型的紧凑度，加大了对应模型的泛化能力，并能有效避免过拟合现象的发生[1]。特

收稿日期：2012-06-16。 作者简介：诸文智(1982—)，男，博士生；张彦斌(通信作者)，男，教授，博士生导师。 基金项目：国家自然科学基金资助项目(61005058)；教育部高等学校博士学科点专项科研基金资助项目(20090201110019，20070698059)。
网络出版时间：2012-12-23 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20121223.1637.007.html

征选择算法以是否将分类器本身作为属性评价量度分为两类：封装类型和过滤类型。封装类型的特征选择算法应用某一分类器算法进行属性子集的评估，并以相对应的分类精度作为选择特征子集的标准[2]。虽然此类算法对特定分类学习过程可搜索属性空间以完成对分类精度的优化，但算法复杂度过高，且其特征选择的结果无法使其他分类学习过程达到最优，所以需要研究与分类器算法选择无关、具有独立属性评价量度的过滤型特征选择算法。

过滤型特征选择算法中常用的属性评价量度有：依赖度[3]、一致性[4]、互信息[5]和距离[6]等。前三类评价量度只对名义型数据有效，对于实际应用中常见的数值型数据，必须离散化后才能处理。由于对数值属性的离散化忽略了属性值与离散标号间的隶属程度，将造成实数空间内邻域结构和序结构上的信息损失[7]。近年来，定义在等价关系上的 Pawlak 粗糙集模型被成功地应用于名义型数据的特征选择中[8]。由于实际数据集多为混合形式，因此提出了模糊粗糙集模型和邻域粗糙集模型这两种扩展粗糙集模型，为将名义型和数值型属性统一在同一系统框架下进行评估提供了理论依据。文献[9]应用高斯核函数获得混合数据对应的模糊矩阵，偏重于对粗糙集模型扩展的研究，并直接以高斯核扩展下的依赖度作为属性评价量度。邻域决策错误率最小算法(NDEM)[10]和邻域软间隔算法[11]在邻域粗糙集模型基础上，提出了各自的类-属性相依性量度作为属性评价量度。NDEM 算法将一致性量度的思想引入属性评价量度的设计，实现了对邻域边界域的分类能力分析；NRS 算法则结合了 Byes 准则和软间隔支持向量机，用于分析属性子空间的分类效果。这两种算法在邻域依赖度的基础上，只考虑了邻域边界域中的多数类的决策信息，无法反映出邻域粒化空间上的整体决策结构信息。

本文在分析邻域信息粒决策结构信息的基础上，从信息论的角度综合分析了邻域粒化空间中决策结构信息与分类复杂性间的关系，进而提出了一种基于邻域决策分辨率的特征选择算法。本算法以保持混合数据决策信息为前提，得到了分类信息更加完整且集合大小更加合理的特征属性子集。

1　基于邻域信息粒的属性量度分析

现实中，数据库多为包含名义型和数值型数据的混合数据库。为了便于混合数据的处理，本文将数据实例表示为决策表形式，定义为

$$\langle U, A_{\mathrm{r}} \cup A_{\mathrm{c}} \cup D, V, f \rangle$$

式中：$U=\{x_1, x_2, \cdots, x_M\}$ 为数据实例的非空有限集合，称为论域；$A_{\mathrm{r}}=\{a_1^{\mathrm{r}}, \cdots, a_l^{\mathrm{r}}, \cdots, a_{n_{\mathrm{r}}}^{\mathrm{r}}\}$ 及 $A_{\mathrm{c}}=\{a_1^{\mathrm{c}}, \cdots, a_l^{\mathrm{c}}, \cdots, a_{n_{\mathrm{c}}}^{\mathrm{c}}\}$ 分别为数值属性集和名义属性集；D 为分类属性，将问题论域划分为互不相交的 s 个集合；$V=\bigcup V_a\ (\forall a \in A_{\mathrm{r}} \cup A_{\mathrm{c}} \cup D)$ 是信息函数 f 的值域；$f=\{f_a: U \to V_a, \forall a \in A_{\mathrm{r}} \cup A_{\mathrm{c}} \cup D\}$ 表示决策表的信息函数，f_a 为属性 a 的信息函数。为了能够合理快速地处理数据，本文统一将数值属性归一化至区间[0,1]。

根据上述定义，邻域粗糙集(NRS)模型可根据数据实例间的关系，生成用于描述论域内其他概念的基本邻域信息粒。针对数值属性集 A_{r} 中任一属性 a_l^{r}，NRS 模型根据明可夫斯基距离量度计算数据实例间的可分辨力，并将可分辨力公式化地表示为距离矩阵

$$\boldsymbol{D}_{\mathrm{N}}^{(l)}(x_i, x_j) = (|\ f(x_i, a_l^{\mathrm{r}}) - f(x_j, a_l^{\mathrm{r}})\ |^P)^{1/P},\quad \forall\, a_l^{\mathrm{r}} \in A_{\mathrm{r}} \tag{1}$$

式中：$\boldsymbol{D}_{\mathrm{N}}^{(l)}(x_i, x_j) \in [0,1]$ 表示了数据实例 x_i 和 x_j 间相对于数值属性 $a_l^{\mathrm{r}}(l=1, \cdots, n_{\mathrm{r}})$ 的可分辨力；$P=2$ 为欧式距离，是实数空间上常用的距离量度。

针对名义属性集 A_{c} 中任一属性 a_k^{c}，NRS 模型根据导出的二元等价关系，可定义类似的距离矩阵

$$\boldsymbol{D}_{\mathrm{E}}^{(k)}(x_i, x_j) = \begin{cases} 0, & f(x_i, a_k^{\mathrm{c}}) = f(x_j, a_k^{\mathrm{c}}), \forall\, a_k^{\mathrm{c}} \in A_{\mathrm{c}} \\ 1, & \text{其他} \end{cases} \tag{2}$$

式中：$\boldsymbol{D}_{\mathrm{E}}^{(k)}(x_i, x_j) \in \{0,1\}$ 表示了数据实例 x_i 和 x_j 间相对于名义属性 $a_k^{\mathrm{c}}(k=1, \cdots, n_{\mathrm{c}})$ 的可分辨力。

从上述距离矩阵定义不难看出，名义属性可以作为数值属性的特殊情况进行处理，从而统一的距离矩阵集 $\boldsymbol{D}_{\mathrm{N}}=\{\boldsymbol{D}_{\mathrm{N}}^{(1)}, \cdots, \boldsymbol{D}_{\mathrm{N}}^{(n_{\mathrm{h}})}\}(n_{\mathrm{h}}=n_{\mathrm{r}}+n_{\mathrm{c}})$ 可用来表示隐藏在属性空间 $A=A_{\mathrm{r}} \cup A_{\mathrm{c}}$ 中的可辨性信息。因此，属性空间 A 对应的距离矩阵 $\boldsymbol{D}_{\mathrm{N}}^{n_{\mathrm{h}}}$ 可通过多个属性对应距离矩阵聚合计算获得

$$[\boldsymbol{D}_{\mathrm{N}}^{n_{\mathrm{h}}}(x_i, x_j)]^2 = \| x_i - x_j \|^2 = \sum_{k=1}^{n_{\mathrm{h}}} [\boldsymbol{D}_{\mathrm{N}}^{(k)}(x_i - x_j)]^2 \tag{3}$$

式中：$\| x_i - x_j \|$ 表示欧式距离量度。

根据上述内容，可定义由属性子集 $B \subseteq A$ 生成的距离矩阵 $\boldsymbol{D}_{\mathrm{N}}^{B}$。在给定了正常数 δ 和属性子集 B 后，实例 $x_i \in U$ 在属性空间 B 上的邻域信息粒为

$$\delta_B(x_i) = \{x_j \mid D_{\mathrm{N}}^{B}(x_i, x_j) < \delta\} \tag{4}$$

实例集合 $\delta_B(x_i)$ 包含了所有在属性空间 B 上与实例 x_i 相似的实例。论域 U 中所有实例对应的

邻域信息粒$\{\delta_B(x_i)\}_{i=1}^n$形成了对论域的一种覆盖，且信息粒间可能存在交叠现象。定义$P(\delta_B(x_i))$表示$\delta_B(x_i)$的概率，$P(\delta_B(x_i),d_j)$表示$\delta_B(x_i)$与类d_j的联合概率，则类d_j相对于$\delta_B(x_i)$的条件概率$P(d_j|\delta_B(x_i))=P(\delta_B(x_i),d_j)/P(\delta_B(x_i))$。当存在$d_j$使$P(d_j|\delta_B(x_i))=1$，表示实例$x_i$属于对应邻域逼近空间中的正域，可以确定的归属于类$d_j$。当$\forall P(d_j|\delta_B(x_i))\neq 1$，且$\sum_{j=1}^{s}P(d_j|\delta(x_i))=1$，表示实例$x_i$属于对应邻域逼近空间中的边界域，无法确定的归属于$s$种类别中的某一类。

邻域依赖度作为最常见的混合属性评价量度，其在对应邻域逼近空间上的定义如下[12]

$$\gamma_B(D)=1-\sum_{i=1}^{n}R_{\mathrm{F}}(\delta(x_i))/n \tag{5}$$

$$R_{\mathrm{F}}(\delta(x_i))=\begin{cases}0, & \exists P(d_j|\delta(x_i))=1,\quad j=1,\cdots,s\\ 1, & \text{其他}\end{cases} \tag{6}$$

式中：n表示论域实例的数量；$R_{\mathrm{F}}(\delta(x_i))$表示邻域信息粒的不可分辨力。从上述定义看出，邻域依赖度只反映了分类信息一致的邻域信息粒所占比率，忽视了邻域边界域内实例对邻域逼近空间整体决策分辨力的贡献。

基于邻域粗糙集模型的特征选择过程实质是通过选取属性子集，完成论域空间的邻域粒化，并使用邻域信息粒逼近决策属性空间。其中如何设计好的属性评价量度，用于指导特征子集的选择，改善对论域空间的粒化和逼近效果，已成为特征选择研究中关键问题之一。

2 基于邻域决策分辨率的特征选择算法(FS-NDDR)

当前基于邻域粗糙集模型的特征选择算法大都直接应用邻域依赖度或提出基于贝叶斯准则属性评价量度，这类算法存在以下不足：在已有分类决策空间上，上述评价量度虽然能够正确表征邻域逼近空间中正域实例对应信息粒的决策信息，但对于边界域实例对应的信息粒则采取了忽略或只考虑其中多数类所提供决策信息的方法，无法准确量度相应属性子集对于决策空间的分辨能力；尤其在不同分类的边界区域，易造成对数据实例的错分，造成学习模型分类能力的下降。因此，本文在量化单个粗糙信息粒决策分辨力的基础上，基于属性子空间的决策分辨率提出一种新的属性评价量度。

2.1 邻域决策确定性定义

对于决策表$\langle U,A\cup D,V,f\rangle$，$B\subseteq A$，常数$\delta\in[0,1]$，$\{d_1,d_2,\cdots,d_s\}$为决策属性$D$所生成的等价类，$\delta(x_i)$表示数据实例$x_i$根据属性子空间$B$相对于常数$\delta$的邻域信息粒，则信息粒$\delta(x_i)$对应的邻域决策确定性表示为

$$N_{\mathrm{c}}(\delta(x_i))=\frac{\left(\sum_{j=1}^{s}(P(d_j\mid\delta(x_i))-1/s)^2\right)^{1/2}}{((s-1)/s)^{1/2}} \tag{7}$$

式中：$P(d_j|\delta(x_i))=\|\delta(x_i)\cap d_j\|/\|\delta(x_i)\|$，$\|*\|$表示集合的模。

$N_{\mathrm{c}}(\delta(x_i))$表征了邻域信息粒$\delta(x_i)$内决策分布的结构信息，信息粒$\delta(x_i)$中的决策分布集中度越高，其中有效分类区分信息蕴含量越高。定义邻域信息粒$\delta(x_i)$决策分布的条件概率为$P^i=(P_1^i,\cdots,P_f^i,\cdots,P_s^i)$，其中$P_f^i=P(d_f|\delta(x_i))$，$f=1,\cdots,s$。当决策分布的条件概率相等时，即$P^i=\left(\frac{1}{s},\frac{1}{s},\cdots,\frac{1}{s}\right)$，$\delta(x_i)$中的决策分布集中度将至最低，其有效分类区分信息蕴含量也最小，相应的N_{c}达到最小值0。相反，当$\delta(x_i)$中实例都属于同一类时，即$\forall P_f^i=1$且$\sum_{f=1}^{s}P_f^i=1$，$\delta(x_i)$中的决策分布集中度最高，其有效分类区分信息蕴含量也最大，相应的N_{c}达到最大值1。随着信息粒$\delta(x_i)$中的决策分布的变化，对应的邻域决策确定性$N_{\mathrm{c}}(\delta(x_i))$的值落在区间[0,1]上。

基于上述分析，对属性子空间B相对于常数δ的邻域信息粒$\delta(x_i)$，邻域决策确定性$N_{\mathrm{c}}(\delta(x_i))$根据$\delta(x_i)$内决策分布与分类均匀分布间偏离的程度，完成了对信息粒分类区分能力的量化。因此，通过融合论域U中所有实例在邻域逼近空间内相应的N_{c}，可作为表征属性子集B对于决策空间D的分辨能力的量度。

2.2 属性评价量度

本文所提出的属性评价量度基于邻域决策确定性N_{c}，在已有邻域逼近空间的上，定义了邻域决策分辨率(NDDR)用于度量混合属性子集B对决策空间D的区分能力，公式如下

$$J_B(D)=1-\sum_{i=1}^{n}[1-N_{\mathrm{c}}(\delta(x_i))]/n \tag{8}$$

式中：n表示论域U中的实例数量。对所提属性评价量度分析如下。

(1)由于评价量度J_B基于邻域粗糙集模型，能

够直接处理名义型数据、数值型数据或混合型数据，所以基于 J_B 的特征选择算法可以避免不必要的信息损失，并获得更加准确的结果。

(2)评价量度 J_B 包含了两部分：针对邻域空间正域元素的下近似逼近依赖度，以及针对邻域空间边界域元素的有效分类信息的量度。相对经典的邻域依赖度，所提评价量度将邻域空间边界域中的分类区分信息引入到算法中，使得评价量度更加全面地反应特征子空间对决策空间的区分能力。

(3)通过改变常数 δ 来调节邻域粒化尺度，从而影响特征子集的选取。这样对于特定的学习过程，总是可以通过设置 δ 的值，来选取合适的邻域粒化尺度，获得更加合理和紧凑的特征选择结果。

(4)评价量度 J_B 能够准确地反应出不同特征子集对决策空间区分能力的差别，且其值域为[0，1]。对属性子集 B 而言，当邻域粒化空间中所有信息粒上的决策分布为分类均匀分布时，J_B 取得最小值 0；相反，当邻域粒化空间中所有信息粒包含的实例都属于某一类时，J_B 取得最大值 1。

2.3　特征选择算法描述

输入：决策表$\langle U, A\cup D, V, f\rangle$，常数 δ 以及算法终止阈值 ε。

步骤 1：初始化特征子集 $R_{\mathrm{ed}}\leftarrow\varnothing$。

步骤 2：对所有属性 $a_i\in A-R_{\mathrm{ed}}$，计算重要性量度 $\mathrm{SIG}(a_i, R_{\mathrm{ed}}, D)=J_{R_{\mathrm{ed}}\cup a_i}(D)-J_{R_{\mathrm{ed}}}(D)$。

步骤 3：选择产生最大 $\mathrm{SIG}(a_k, R_{\mathrm{ed}}, D)$ 值的对应属性 a_k。

步骤 4：判断当前特征子集是否满足 $\mathrm{SIG}(a_k, R_{\mathrm{ed}}, D)>\varepsilon$，满足则更新特征子集 $R_{\mathrm{ed}}\leftarrow R_{\mathrm{ed}}\cup a_k$，转步骤 5；不满足则算法结束。

步骤 5：判断当前特征子集是否满足 $R_{\mathrm{ed}}\neq A$，满足则转步骤 2；不满足则算法结束。

输出：混合属性 A 相对于决策属性 D 的特征子集 R_{ed}。

上述算法中有两个关键步骤：一是确定数据实例的邻域信息粒；二是根据邻域粒化结果，计算邻域决策分辨率。对于每一属性而言，确定所有数据实例的邻域信息粒的时间复杂度为 $O(n\log n)$。通过聚合单一属性空间对应的实例间的距离矩阵，进而获得任意属性子集对应的邻域粒化空间。对任一邻域信息粒而言，计算 N_{c} 的时间复杂度为 $O(n)$。因此，本文所提特征选择算法的时间复杂度为 $O\left(\frac{(2N-K)(K+1)}{2}\right)(n\log n+n)$，其中 N 为原始集 A 的模，K 为结果集 R_{ed} 的模。

3　实验结果及分析

3.1　仿真实验

为比较所提邻域决策分辨率在不同粒化等级上的有效性，选择了 6 个加州大学欧文分校(UCI)机器学习数据集[13]进行实验比较，所选数据集如表 1 所示。为了统一起见，所选数据集中所有数值型属性的值都线性归一化至区间[0，1]，而所有名义型属性的值都编码为一系列互异的值。δ 的取值以 0.01 为步长从 0.05 到 0.4 变化。实验中选择线性支持向量机(Linear SVM)和基于径向基函数的支持向量机(RBF SVM)作为分类学习算法，且其参数依照文献[14]方法进行优化调节，实验采用十折交叉检验法。所得各数据集在不同 δ 下的 J_B 与对应分类精度的相关系数如图 1 所示。

表 1　数据集相关信息

编号	数据集	实例数量	属性数量	分类数量
1	Glass	214	9	6
2	Heart	270	13	2
3	Sonar	208	60	2
4	Wdbc	569	30	2
5	Wine	178	13	3
6	Vehicle	846	18	4

从图 1 中不难发现，J_B 随邻域粒化等级的变化较为稳定，且能在很大的 δ 取值范围内获得对相关属性较好的分类评估能力。根据前文分析，δ 取值越大，每一信息粒中所包含的实例越多，相应邻域近似空间中的粒化结果就越粗糙，这将影响量度 J_B 评估属性子集分类能力的有效性。结合图 1 中的 6 幅子图可发现，当 δ 的取值增加到一定程度时，计算所得相关系数开始递减。大致上说，当 δ 在[0.06，0.18]之间取值时，量度 J_B 可同时在 Linear SVM 和 RBF SVM 上获得较高的分类评估效果。应用 J_B 进行特征选择，δ 在上述候选区间内取值时发现：当 δ 在[0.06，0.08]之间取值时，特征选择算法过早收敛，且无法在实验中获得足够的特征属性；当 δ 在[0.15，0.18]之间取值时，特征选择算法在实验中获得过多的特征属性，用于对决策属性的近似。因此，δ 不宜在这些区间上取值，相对而言，δ 在[0.09，0.14]之间取值较为理想。在本文后续实验中，邻域 δ 设置为 0.1。

为了验证所提算法的性能，选择了以下 5 种特

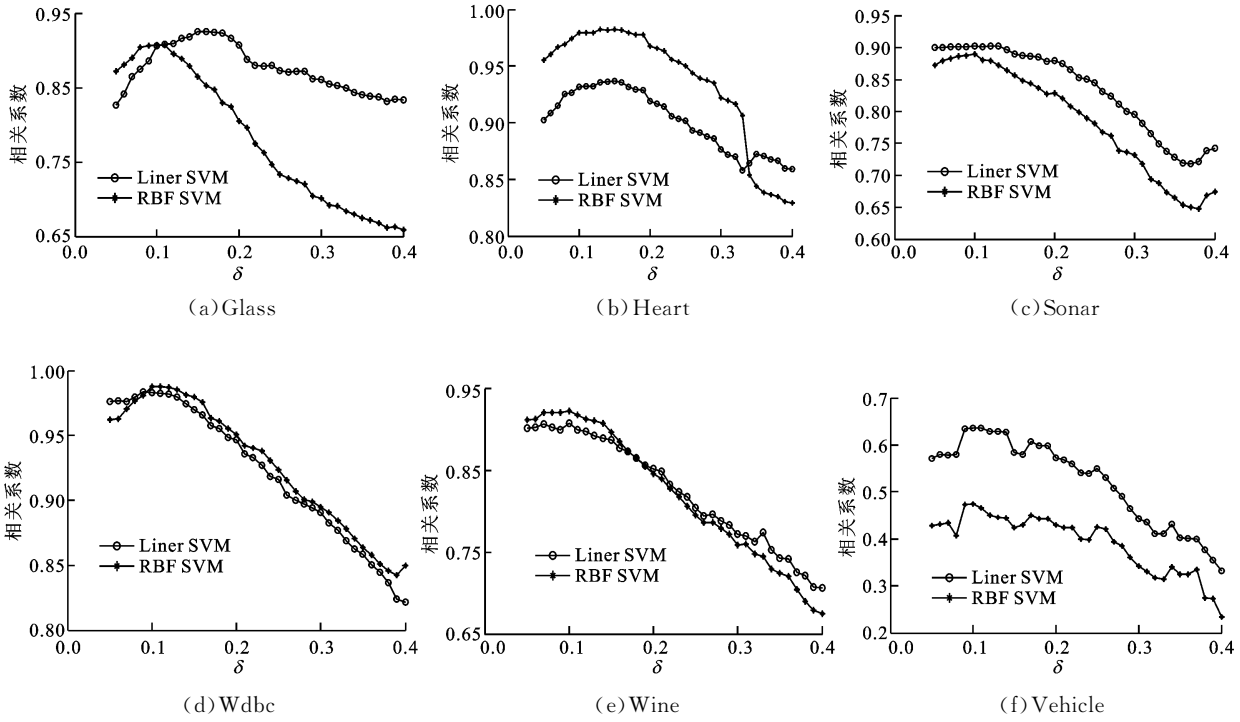

图 1 2 种分类方法在不同 δ 下的相关系数

征选择算法进行实验比较:基于邻域决策分辨率(FS-NDDR)、基于邻域辨识率(FS-NRR)[10]、基于高斯核依赖度(FS-GD)[9]和基于邻域依赖度(FS-ND)[12]的特征选择方法(这 4 种方法无需离散化处理,可直接应用于混合数据),以及经最小描述长度方法离散化[15]后基于 Pawlak 粗糙集模型的特征选择算法[8](MDL-RS)。仿真实验在 3 个人造数据集和 8 个 UCI 机器学习数据集上进行,并应用分类决策问题中常用的性能评估方法 C5.0[16]和 RBF SVM 进行分类实验。FS-GD 算法中核参数设置为 0.1[9],而 FS-NRR 与 FS-ND 算法中的邻域 δ 则设置为 0.14[10]。分类学习过程中,实验采用十折交叉检验法进行验证。

为检验提出算法的稳定性,将 5 种算法应用于 3 个人造噪声数据集上,并比较实验结果。首先,根据文献[17]的数据构造模型以及均匀分布函数,生成一个两分类数据库,共 500 个实例,40 个数值属性。属性集中 5 个属性包含于构造模型,10 个属性与模型所包含属性具有线性组合关系,余下的属性与构造模型无关。在此两分类数据库基础上,通过人为随机错分 5%、10%的数据实例对应的类标,并对每一属性加入标准差为 0.2σ 的高斯噪声的方法,建立噪声数据集,其中 σ 为原始人造数据集中所有单一属性对应标准差的最大值。对上述 3 个人造噪声数据集,应用 5 种特征选择算法进行属性选择,并以 RBF SVM 分类器对属性选择后的数据集进行交叉验证,获得不同算法在 3 个人造噪声数据集上对应的分类精度 A_{cc},如表 2 所示。

从表 2 可以看出,FS-NDDR、FS-NRR 和 FS-GD 算法在任一噪声数据集上进行选择的结果集合,虽然在特征数量上完全相同,但特征子集间互不相同。由于邻域依赖度(ND)对 3 个噪声数据任意单一属性的量度值为 0,因此 FS-ND 在算法第一轮循环就跳出,其特征选择结果为空集。对于 MDL-RS 算法而言,由于经过离散化后的噪声数据分类信息损失严重,造成其在 3 个人工数据集上的特征选

表 2 噪声数据上特征数量和分类精度结果

噪声水平/%	FS-NDDR		FS-NRR		FS-GD	
	特征数量	A_{cc}/%	特征数量	A_{cc}/%	特征数量	A_{cc}/%
0	4	93.16±4.01	4	92.57±3.92	4	92.48±4.08
5	5	89.77±4.62	5	89.25±4.93	5	88.46±4.95
10	5	83.21±5.31	5	82.71±5.53	5	81.87±6.26

择结果也为空集。因此，将各算法所选特征子集对应的分类精度作为比较算法稳定性的主要依据，且分类精度越高，相应算法稳定性就越强。相比其他算法，FS-NDDR 算法选择的特征子集获得了更高的分类精度，说明算法具有较好的稳定性，能够更加有效地处理带噪声的数据。

对表 3 所示数据集应用 5 种特征选择算法，各算法所选择属性集合对应的特征数量如表 4 所示，且各算法在相应数据集上选择的特征子集互不相同，加粗数据为各数据集上相应算法对应特征数量的最小值。以 C5.0 和 RBF SVM 对各算法对应的特征选择结果进行分类学习，相应分类性能结果如表 5、表 6 所示，加粗数据为各数据集上相应算法对应分类精度的最大值。

表 3　数据库相关信息

编号	数据库	实例数量	名义属性数量	连续属性数量	类别数量
1	Dermatology	366	33	1	6
2	German	1 000	13	7	2
3	Hepatitis	155	13	6	2
4	Heart	270	7	6	2
5	Sonar	208	0	60	3
6	Soybean-large	683	35	0	19
7	Thyroid	3 772	15	6	3
8	Vehicle	846	0	18	4

表 4　各算法所选属性集对应的特征数量

数据集	原始数据	FS-NDDR	FS-NRR	FS-GD	FS-ND	MDL-RS
Dermatology	34	**10**	11	**10**	**10**	**10**
German	20	9	**8**	9	10	
Hepatitis	19	**6**	8	**6**	7	10
Heart	13	**7**	8	**7**	8	
Sonar	60	**5**	**5**	6	6	
Soybean-large	35	**11**	13	**11**	13	**11**
Thyroid	21	**5**	**5**	8	11	6
Vehicle	18	**9**	**9**	**9**	**9**	**9**

表 5　基于 C5.0 的分类精度　%

数据集	原始数据	FS-NDDR	FS-NRR	FS-GD	FS-ND	MDL-RS
Dermatology	93.89±1.33	**93.96±0.86**	93.67±1.43	92.56±2.21	93.32±0.83	93.62±1.09
German	71.57±1.24	**72.84±1.02**	71.81±1.04	72.33±1.05	72.78±1.13	
Hepatitis	82.55±3.53	**81.13±3.14**	79.97±2.76	78.74±3.94	80.07±2.34	78.23±2.88
Heart	77.12±2.28	**79.88±2.33**	79.61±2.29	78.09±2.47	79.51±2.44	
Sonar	73.48±3.55	**75.96±2.43**	75.58±2.96	74.50±2.78	73.41±3.05	
Soybean-large	92.42±1.12	**87.97±0.92**	87.55±1.11	87.41±1.12	86.56±1.07	80.18±1.54
Thyroid	99.70±0.12	98.20±0.22	98.17±0.26	**99.57±0.16**	94.45±0.36	97.98±0.14
Vehicle	72.93±1.26	**70.98±1.04**	66.91±0.96	69.78±1.33	69.05±1.88	65.14±1.14

表 6　基于 RBF SVM 的分类精度　%

数据集	原始数据	FS-NDDR	FS-NRR	FS-GD	FS-ND	MDL-RS
Dermatology	32.80±2.47	**91.72±5.11**	75.84±7.03	88.55±5.22	83.26±5.71	83.26±5.70
German	70.39±1.05	71.16±2.92	71.44±2.83	72.17±3.11	**73.55±3.36**	
Hepatitis	80.03±8.94	**85.24±6.96**	84.35±8.34	82.59±6.61	85.09±8.29	80.89±7.25
Heart	80.56±7.09	**82.54±6.54**	79.66±6.81	81.36±7.18	82.44±6.39	
Sonar	87.28±8.24	77.94±9.28	**79.54±7.66**	77.76±9.25	77.04±9.66	
Soybean-large	45.42±4.66	**69.77±4.67**	53.69±4.78	69.33±4.48	58.51±4.87	52.00±4.64
Thyroid	93.55±0.44	**93.72±0.49**	93.66±0.43	93.57±0.46	92.88±0.44	93.67±0.41
Vehicle	75.26±3.47	**71.13±3.98**	70.68±3.96	70.63±3.77	70.17±3.74	68.55±4.19

整个仿真实验流程在个人计算机上完成，其主要配置为：双核 1.8 GHz 处理器和 2.5 GB 内存。各算法在不同数据集上的有效运行时间如表 7 所示，加粗数据为各数据集上相应算法对应有效运行时间的最小值。从有效运行时间的数据统计来说，本文算法的执行效率高于其他 4 种特征选择算法。

表 7 各特征选择算法的有效运行时间 s

数据集	FS-NDDR	FS-NRR	FS-GD	FS-ND	MDL-RS
Dermatology	**5.47**	5.49	59.23	12.01	12.11
German	**17.91**	19.13	212.61	41.43	
Hepatitis	**0.37**	0.44	3.12	0.65	0.72
Heart	**0.81**	0.87	9.16	1.45	
Sonar	**1.52**	1.61	28.59	3.11	
Soybean-large	**12.58**	15.21	219.34	59.91	60.46
Thyroid	**160.65**	244.48	758.67	506.35	321.04
Vehicle	**11.06**	12.43	128.84	22.17	26.61

对表 4～表 6 的实验结果分析如下。

与原始数据相比，本文算法虽然大幅降低了特征数量，但相应特征子集对应的分类精度没有出现明显的下降，反而在相当一部分数据集上获得了提升。这一现象说明 FS-NDDR 算法能够有效对原始数据中不相关和冗余的属性进行约简，保留其中相关性较强的特征属性。

由上节分析可知，NDDR 不仅反映了邻域空间中正域元素的下近似逼近依赖度，而且基于邻域粒化空间中决策结构信息，完成了对邻域边界域相应决策分辨力的度量。综合各特征选择算法在两个分类器上的结果不难发现：由于 NDDR 相比 NRR 可更加有效地获得邻域边界域中的决策信息，因此 FS-NDDR 算法可在特征数量相当的情况下，得到分类性能更佳的特征子集。FS-GD 与 FS-ND 算法的属性评价量度中仅仅反映了相应粒化空间中正域的决策信息，因此与 FS-NDDR 相比，FS-ND 所选特征子集的特征数量更大，对应分类性能更差。从实验结果上看，虽然 FS-ND 比 FS-ND 有了相当的改善，但 FS-NDDR 算法依然在绝大多数数据集上获得了更加优良的结果。对于 MDL-RS 算法，由于离散化后的数据集损失了部分分类信息，造成其在数据集 German、Heart 和 Sonar 上特征选择的结果为空集。在剩下的数据集上，相比 MDL-RS 算法所得特征子集，FS-NDDR 算法以特征数量相当或更小的特征子集，获得了更高的分类精度。总的来说，与其他 4 种特征选择算法相比，FS-NDDR 算法可更加有效地处理带噪声的数据，并在提高分类精度的基础上，选择特征数量相对较小的特征子集。

3.2 磨矿浓度预测结果

磨矿浓度的在线检测是选矿厂磨矿分级系统精确控制得以有效实施的关键，但在实际工业中，由于现场环境恶劣、影响因素众多等原因一直未能解决[18]。结合大量先验知识，本文选取选矿厂磨矿分级系统工业数据库中给矿量 M、返砂水量 W_{rs}、排矿水量 W_{ca}、电耳电流 I_{ee}、磨机电流 I_m 及分级机电流 I_c 等 6 个属性组成预选属性集，以磨矿浓度作为决策属性，选取磨机内矿浆处于高浓度（质量分数为 82%）、正常浓度（质量分数为 80%）及低浓度（质量分数为 79%）等稳定工况时的 900 组数据构成磨矿浓度决策表。应用上述 3 种性能相近且可直接处理数值数据的特征选择方法 FS-NDDR、FS-NRR、FS-GD，各自选择出关联性较强的特征属性作为磨矿浓度实时预测的辅助变量，如表 8 所示。

表 8 磨矿浓度预测辅助变量

算法	特征数量	特征子集
FS-NDDR	4	M, I_{ee}, I_m, W_{rs}
FS-NRR	5	$M, I_{ee}, I_m, W_{ca}, I_c$
FS-GD	5	$M, W_{rs}, I_m, W_{ca}, I_c$

考虑到预测模型在鲁棒性和适应性上的要求，本文选择基于语言控制规则和模糊逻辑推理的模糊系统，以各算法所选特征子集作为模型输入变量，以磨矿浓度作为输出变量，构造各自对应的磨矿浓度模糊预测模型。应用遗传算法对训练集进行学习，获取各算法对应的模糊语义规则[19]，并结合高斯模糊器、乘积推理机和中心平均解模糊器，对验证集进行稳定工况和磨矿浓度值的估计，实验采用十折交叉检验法，结果如表 9 所示。

表 9　磨矿浓度预测实验结果

算法	工况分类正确率/%	训练相对误差/%		检验相对误差/%	
		平均值	最大值	平均值	最大值
FS-NDDR	98.34	0.19	0.24	0.59	0.88
FS-NRR	97.93	0.20	0.28	0.63	0.90
FS-GD	97.57	0.22	0.30	0.70	1.01

由表 9 易知，根据 FS-NDDR 算法所选特征子集获得了最高的工况分类正确率，并使相应的磨矿浓度预测模型在三种不同工况下，得到了较另两种算法更好的预测效果。综合上述实验结果，本文所提特征选择算法 FS-NDDR 的效能优于 FS-NRR 算法、FS-GD 算法、FS-ND 算法和 MDL-RS 算法。

4　结　论

本文提出了一种基于邻域决策分辨率的特征选择算法——FS-NDDR。该方法充分考虑了属性子集对决策空间的分辨能力与对应邻域粒化空间中决策分布结构信息的关系，在合理化约简特征数量的基础上，选择具有最大决策分类能力的特征子集，提高了约简后数据集的分类精度。仿真和实际应用的结果表明，FS-NDDR 算法的性能优于目前主流基于邻域粗糙集的特征选择方法，但其在高维数据库中的有效运行时间有所增加。后续工作可以通过优化算法流程来提升 FS-NDDR 算法的执行效率。

参考文献：

[1] LIU H, YU L. Toward integrating feature selection algorithms for classification and clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 491-502.

[2] GUYON I, ELISSEEFF A. An introduction to variable and feature selection [J]. The Journal of Machine Learning Research, 2003, 3(7/8): 1157-1182.

[3] MITRA P, MURTHY C, PAL S. Unsupervised feature selection using feature similarity [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312.

[4] DASH M, LIU H. Consistency-based search in feature selection [J]. Artificial Intelligence, 2003, 151(1/2): 155-176.

[5] DASH M, CHOI K, SCHEUERMANN P, et al. Feature selection for clustering: a filter solution [C]// Second IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2002: 115-122.

[6] HO T, BASU M. Complexity measures of supervised classification problems [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 289-300.

[7] CHING J, WONG A, CHAN K. Class-dependent discretization for inductive learning from continuous and mixed-mode data [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(7): 641-651.

[8] POLKOWSKI L. Rough sets: mathematical foundations[M]. New York, USA: Physica-Verlag, 2002: 1-36.

[9] HU Q, ZHANG L, CHEN D, et al. Gaussian kernel based fuzzy rough sets: model, uncertainty measures and applications [J]. International Journal of Approximate Reasoning, 2010, 51(4): 453-471.

[10] HU Q, PEDRYCZ W, YU D, et al. Selecting discrete and continuous features based on neighborhood decision error minimization [J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B Cybernetics, 2010, 40(1): 137-150.

[11] HU Q, CHE X, ZHANG L, et al. Feature evaluation and selection based on neighborhood soft margin [J]. Neurocomputing, 2010, 73(10/11/12): 2114-2124.

[12] HU Q, YU D, LIU J, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178(18): 3577-3594.

[13] FRANK A, ASUNCION A. UCI machine learning repository [DB/OL]. [2010-08-22]. http://archive.ics.uci.edu/ml.

[14] WANG J, WU X, ZHANG C. Support vector machines based on k-means clustering for real-time business intelligence systems [J]. International Journal of Business Intelligence and Data Mining, 2005, 1(1): 54-64.

[15] KWAK N, CHOI C. Input feature selection for classification problems [J]. IEEE Transactions on Neural Networks, 2002, 13(1): 143-159.

[16] QUINLAN R. Data mining tools See5 and C5.0 [EB/OL]. [2012-02-20]. http://www.rulequest.com/see5-info.html.

[17] GHEYAS I, SMITH L. Feature subset selection in large dimensionality domains [J]. Pattern Recognition, 2010, 43(1): 5-13.

[18] 张晓东. 先进控制技术在选矿过程控制中的应用研究 [D]. 沈阳：东北大学，1999.

[19] KARR C L, WECK B. Computer modeling of mineral processing equipment using fuzzy mathematics [J]. Minerals Engineering, 1996, 9(2): 183-194.

（编辑　杜秀杰）