计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190100196

基于距离比值尺度的模糊粗糙集属性约简

陈毅宁[1] 陈红梅[2]
1 西南交通大学信息科学与技术学院 成都 611756
2 西南交通大学云计算与智能技术高校重点实验室 成都 611756
(953837210@qq.com)

摘 要 属性约简能有效地去除不必要属性,提高分类器的性能。模糊粗糙集是处理不确定信息的重要范式,能有效地应用于属性约简。在模糊粗糙集中,样本分布的不确定性会影响对象的近似集,进而影响有效属性约简的获取。为有效地定义近似集,文中提出了基于距离比值尺度的模糊粗糙集,该模型引入了基于距离比值尺度的样本集的定义,通过对距离比值尺度的控制,避免了样本分布不确定性对近似集的影响;给出了该模型的基本性质,定义了新的依赖度函数,进而设计了属性约简算法;以SVM,NaiveBayes和J48作为测试分类器,在UCI数据集上评测所提算法的性能。实验结果表明,所提出的属性约简算法能够有效获取约简并提高分类的精度。

关键词:属性约简;模糊粗糙集;距离比值尺度
中图法分类号 TP301.6

Attribute Reduction of Fuzzy Rough Set Based on Distance Ratio Scale

CHEN Yi-ning[1] and CHEN Hong-mei[2]
1 School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China
2 Key Laboratory of Cloud Computing and Intelligent Technology, Southwest Jiaotong University, Chengdu 611756, China

Abstract Attribute reduction can effectively remove the unnecessary attributes in order to improve the performance of the classifiers. Fuzzy rough set theory is an important formal of processing the uncertain information. In the fuzzy rough set model, the approximations of an object may be affected by uncertain distribution of samples. Consequently, acquiring effective attribute reduction may be influenced. In order to effectively define approximations, this paper proposed a novel fuzzy rough set model named distance ratio scale based fuzzy rough set. The definition of samples based on distance ratio scale is introduced. The influence of uncertain distribution of samples to approximations is avoided by controlling the distance ratio scale. The basic properties of this fuzzy rough set model are presented and the new dependent function is defined. Furthermore, the algorithm for attribute reduction is designed. SVM, NaiveBayes, and J48 were used as test classifier executed on UCI data sets to verify the performance of the proposed algorithm. The experimental results show that attribute reduction can be effectively obtained by the proposed attribute reduction algorithm and the classification precisions of classifiers are improved.

Keywords Attribute reduction, Fuzzy rough set, Distance ratio scale

1 引言

高维数据普遍存在于医疗诊断、图像标注、文本挖掘等领域[1-3]。一般而言,对于特定的分类学习任务,其中的部分特征可能是冗余或不相关的。冗余特征会使计算量过高,不相关的特征会降低分类学习算法的分类能力。因此,选出相关的、必要的属性具有重要意义。

粗糙集是由Pawlak于1982年提出的一种分析不确定性知识的数学工具[4]。粗糙集理论为数据挖掘提供了一种有效处理冗余数据的手段。在保证数据分类能力不变的条件下,通过知识约简去除其中的冗余信息,导出分类规则或决策。该理论已被成功应用于特征选取、规则学习、模式识别等领域[5-8]。然而,经典粗糙集不能有效地处理数值数据。针对这个问题,学者们提出了邻域粗糙集和模糊粗糙集的概念[9-10]。其中,模糊粗糙集利用其模糊性和粗糙性来处理连续的特征,而无须离散数值属性。近年来,在模糊粗糙集中,学者们提出了许多重要的约简算法。例如,Jensen等用模糊粗糙依赖度函数来测量属性的质量,并提出了一种快速约简的特征选择算法[11-12];Hu等将信息熵引入到模糊粗糙集中,并用此来度量条件属性与决策属性之间的依赖度[13];Wang等提出了基

到稿日期:2019-01-23 返修日期:2019-06-05 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61572406)
This work was supported by the National Natural Science Foundation of China (61572406).
通信作者:陈红梅(hmchen@swjtu.edu.cn)

于距离测度的模糊粗糙集中的属性约简算法[14]。

近似算子是粗糙集的核心内容，如何有效地定义近似算子是进一步进行知识发现的基础。针对数据分布的不确定性对模糊粗糙集中近似集的影响，本文提出了基于距离比值尺度的模糊粗糙集，定义了新的模糊依赖度函数，使用前向贪心算法来获得约简的属性集合，并将所提算法与其他算法进行对比实验，实验结果显示了该约简算法是可行且有效的。

2 模糊粗糙集

模糊粗糙集可有效地分析数值属性和符号属性同时存在时信息系统的不确定性。该理论主要由粗糙集理论和模糊集理论结合而生，具有粗糙逼近和模糊粒化的特征，从而能够直接对数值属性进行处理。

定义 1[15]（决策信息系统） 一个决策信息系统定义为 $DIS=\langle U,C\cup D,V,f\rangle$，其中论域 U 为非空有限的对象集合；C 为条件属性集，D 为决策属性集，$C\cap D=\varnothing$；$V=\bigcup_{a\in A}V_a$ 称为值域，V_a 表示属性 a 的值域；f 表示 $U\times A\to V$ 的映射函数，$\forall a\in A,x\in U,f(x,a)\in V_a$。

定义 2[16]（模糊等价关系） 已知论域 U 为有限非空集合，论域 U 上的模糊关系 R 称为模糊等价关系，如果 R 满足：

(1) $R(x,y)=1$（自反性）；

(2) $R(x,y)=R(y,x)$（对称性）；

(3) $R(x,y)\geqslant\sup\limits_{z\in U}\min\{R(x,z),R(z,y)\}$（传递性）。

定义 3[9]（模糊粗糙集） 已知论域 U 是有限非空集合，R 是 U 上的模糊等价关系，$F(U)$ 是 U 上的模糊幂集，则模糊集 $A\in F(U)$ 的下近似集 $\underline{RA}(x)$、上近似集 $\overline{RA}(x)$ 分别定义为：

$$\begin{cases}\underline{RA}(x)=\inf_{y\in U}\max\{1-R(x,y),A(y)\}\\ \overline{RA}(x)=\sup\limits_{y\in U}\min\{R(x,y),A(y)\}\end{cases}\tag{1}$$

定义 4[17]（T-模） 记单元区间 $I=[0,1]$，有二元映射函数 T：若 $T:I\times I\to I,a,b,c,d\in I$ 满足下列条件：

(1) $T(a,1)=a$（两极律）；

(2) $T(a,b)=T(b,a)$（交换律）；

(3) $T(T(a,b),c)=T(a,T(b,c))$（结合律）；

(4) $a\leqslant c,b\leqslant d\Rightarrow T(a,b)\leqslant T(b,a)$（单调律）。

则称 T 为 I 上的三角模（或 T-模）；若三角模 T 满足 $T(a,0)=a$，则称其为 S-模（或 T-余模）。此外，补算子 N 是一个从 $[0,1]\to[0,1]$ 的单减映射，且满足 $N(0)=1$ 和 $N(1)=0$，其中 $N_s(a)=1-a$ 是标准的补算子。

定义 5[18]（基于 T-模的模糊粗糙集） 设 U 是一个非空论域，R 是 U 上的任意模糊关系，$\forall A\in F(U)$，则下近似集和上近似集分别定义为：

$$\begin{cases}\underline{R_{\mathcal{I}}}A(x)=\inf\limits_{y\in U}\mathcal{I}(R(x,y),A(y))\\ \overline{R_T}A(x)=\sup\limits_{y\in U}T(R(x,y),A(y))\end{cases}\tag{2}$$

定义 6[19]（核函数） 设 U 是一个有限论域，如果实值函数 $k:U\times U\to R$ 满足条件：$\forall x,y\in U$，有 $k(x,y)=k(y,x)$，且半正定，则称 k 为核函数。其中，高斯核函数为 $k_G(x,y)=\exp(-\|x-y\|^2/2\delta^2)$，这里 $\|\cdot\|$ 是 x 和 y 之间的欧氏距离。

定义 7[20]（核模糊粗糙集） 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，$R_G$ 是论域 U 上的高斯核模糊等价关系，模糊集 $A\in F(U)$，则基于高斯核函数的模糊粗糙集下近似集和上近似集分别定义为：

$$\begin{cases}\underline{R_G}A(x)=\inf\limits_{y\in U}\mathcal{I}_{T_{\cos}}(R_G(x,y),A(y))\\ \overline{R_G}A(x)=\sup\limits_{y\in U}T_{\cos}(R_G(x,y),A(y))\end{cases}\tag{3}$$

3 一种基于距离比值的模糊粗糙集

本节介绍经典模糊粗糙集中近似算子存在的问题，进而给出改进的模糊粗糙集中的相关概念。

3.1 问题的提出

由经典模糊粗糙集的定义可知，样本隶属于它所在类的确定性程度是由与其异类的样本集合中的最近邻的距离来度量的。若该最近邻样本偏离其所属类的程度较大，则极有可能导致该样本隶属于它所在类的下近似值过小；若该样本所在类的大多数样本均选择该最近邻进行下近似相关的计算，则会导致整体的属性依赖度过低，从而得出错误的属性约简结果。因此，如何选取合适的样本进行下近似相关的计算成为了一个关键问题。

例如，当样本分布情况如图 1 所示时，样本 y_1 偏离 Class2 的程度较大，根据模糊粗糙集的定义，仍然会选择样本 y_1 与样本 x 之间的距离来度量 x 隶属于 Class1 的确定性程度。显然，在此例中，选择样本 y_2 进行下近似相关计算更为妥当。

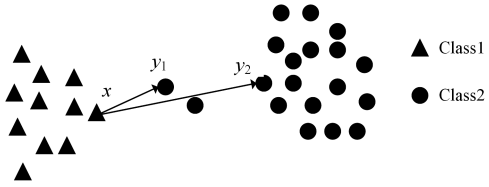


图 1 样本分布图示

Fig. 1 Sample distribution

3.2 基于距离比值尺度的样本集

本文采用高斯核函数 $R_G(x,y)$ 来计算样本之间的相似性：

$$R_G(x,y)=\exp\left(-\frac{\|x-y\|^2}{\sigma}\right)\tag{4}$$

则样本 x 与样本 y 之间的伪距离可表示为[20]：

$$d(x,y)=1-R_G(x,y)\tag{5}$$

定义 8（距离比值） 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，$R_G$ 是论域 U 上的高斯核模糊等价关系，论域 U 被决策属性 D 划分为 $\{D_1,D_2,\cdots,D_r\},x\in D_i$，$\vec{D_x}=(d_1,d_2,\cdots,d_i,\cdots,d_m)$ $(0\leqslant i<m,d_i\leqslant d_{i+1}$，其中，$d_i=d(x,y_i')$，$y_i'\notin D_i$，$y_i'$ 为 d_i 所对应的样本）为距离 x 的 m 个样本的距离有序向量。距离比值 r_i 定义为 $r_i=d_i/d_{i+1}$，其中 $1\leqslant i<m$，当 $d_{i+1}=0$ 时，令 $r_i=1$。

定义 9（基于距离比值尺度的样本集） 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，$R_G$ 是论域 U 上的高斯核模糊等价关系，论域 U 被决策属性 D 划分为 $\{D_1,D_2,\cdots,D_r\},x\in D_i$，则当计算 x 隶属于 D_i 的隶属度时，令 λ 表示距离比值尺度，$\vec{D_x}$ 对应的距离比值向量 $\vec{DR_x}=(r_1,r_2,\cdots,r_k)(k<m)$，$r_i$ 对应的对象标识为 y'_{i+1}，则基于距离比值尺度的样本集 *properY* 定义为：

$properY=\{y'_{i+1}\mid r_i<\lambda, y'_{i+1}\in D_j, x\notin D_j, D_j\in U/D\}$ (6)

若 $properY$ 为空，则令 $properY=\{y_1'\}$。

定理 1 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，若给定距离比值尺度 λ_1 和 $\lambda_2(\lambda_1<\lambda_2)$，$properY_1$ 为 λ_1 对应的基于距离比值尺度的样本集，$properY_2$ 为 λ_2 对应的基于距离比值尺度的样本集，则 $|properY_1|\leqslant|properY_2|$。

证明：由定义 9 可证。

由定理 1 可知，基于距离比值尺度的样本集 $properY$ 的势会随着距离比值尺度 λ 的变化而变化，如果将下近似的隶属度的选择限定在 $properY$ 集合中，则能够有效地避免偏离样本对下近似算子的影响。此外，偏离样本集合的样本数量一般较少，因此 k 的取值为较小的整数。

下文给出一个例子来说明以上的定义。

例 1 给定样本 $x\in D_1$，与 x 异类的样本集合为 $Y=\{y_1,y_2,\cdots,y_{10}\}$，则有 $\vec{D}_x=(d_1,d_2,\cdots,d_{10})=(0.06,0.12,0.13,0.34,0.35,0.35,0.36,0.37,0.38,0.39)$，取 $k=6$，则 $\vec{D}_x$ 对应的距离比值向量 $\overrightarrow{DR}_x=(r_1,r_2,\cdots,r_6)=(0.5,0.92,0.38,0.97,1,0.97)$，若距离比例尺度 λ 的取值为 0.6，则基于距离比值尺度的样本集 $properY=\{y_2',y_4'\}$。若 λ 的取值为 0.4，则此时样本集 $properY=\{y_4'\}$。若 λ 的取值过小，如小于 0.38 时，样本集 $properY=\{y_1'\}$。由此可知，λ 的取值大小控制着样本间隶属度的差异情况，λ 越小，满足条件的样本间的差异性越大。随着 λ 的增大，逐渐降低满足条件的样本间的差异性。

3.3 基于距离比值尺度的模糊粗糙集

由距离比值尺度的样本集的定义可知，距离比值尺度的样本集 $properY$ 中的样本为关键样本。基于此，给出积极样本和消极样本的定义。其中，$properY$ 中的积极样本表示该样本后的所有样本偏离该样本所属类的程度均较小，而 $properY$ 中的消极样本表示该样本前的所有样本的偏离程度较大。

定义 10（积极样本和消极样本） 给定样本集合 $X=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$，其中 i 表示样本 x_i 加入集合 X 的顺序，将第一个和最后一个加入集合中的样本分别定义为积极样本 X_{first} 和消极样本 X_{last}，有：

$$\begin{cases}X_{\text{first}}=x_1\\X_{\text{last}}=x_n\end{cases}\tag{7}$$

仍然以例 1 进行说明，当 λ 的取值为 0.6 时，$poperY_{\text{first}}=y_2'$ 表示该样本为合适的积极样本，忽视的样本为 y_1'；而 $poperY_{\text{last}}=y_4'$ 表示该样本为消极样本，认为该样本前的所有样本的偏离程度较大，忽略的样本为 y_1',y_2',y_3'。根据积极样本和消极样本的定义，我们给出基于距离比值尺度的模糊粗糙集的相关定义。

定义 11（基于距离比值尺度的模糊粗糙集） 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，其中非空有限论域 U 为 $U=\{x_1,x_2,\cdots,x_n\}$，C 为条件属性的集合，D 为决策属性。论域 U 在决策属性 D 下的划分为 $U/D=\{D_1,D_2,\cdots,D_r\}$，则距离比值尺度 λ 下基于积极样本或消极样本的模糊粗糙集的下上近似可分别定义为：

$$\begin{cases}\underline{R_G^{\lambda}}(D_i)(x)_F=\sqrt{1-R_G^2(x,properY_{\text{first}})}\\\overline{R_G^{\lambda}}(D_i)(x)_F=R_G(x,properY_{\text{first}})\end{cases}\tag{8}$$

$$\begin{cases}\underline{R_G^{\lambda}}(D_i)(x)_L=\sqrt{1-R_G^2(x,properY_{\text{last}})}\\\overline{R_G^{\lambda}}(D_i)(x)_L=R_G(x,properY_{\text{last}})\end{cases}\tag{9}$$

为了方便描述，下文中的定义均是基于消极样本定义的，基于积极样本的定义类似，不再赘述。

定义 12（属性依赖度） 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，其中非空有限论域 U 为 $U=\{x_1,x_2,\cdots,x_n\}$，C 为条件属性集，D 为决策属性集。对于属性子集 $B\subseteq C$，样本 $x\in U$ 隶属于决策 D 关于 B 的正域的隶属度为：

$$POS_B^{\lambda}(D)(x)=\sup_{X\in U/D}\underline{R_{G_B}^{\lambda}}(X)(x)\tag{10}$$

属性依赖度定义为：

$$\gamma_B^{\lambda}=\frac{\sum_{x\in U}POS_B(D)(x)}{|U|}\tag{11}$$

性质 1 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，$0<\lambda<1$，对于属性子集 $B_1\subseteq B_2\subseteq C$，有 $POS_{B_1}^{\lambda}(D)\subseteq POS_{B_2}^{\lambda}(D)$。

性质 2 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，$0<\lambda_1<\lambda_2<1$，对于属性子集 $B\subseteq C$，有 $POS_B^{\lambda_1}(D)\subseteq POS_B^{\lambda_2}(D)$。

定义 13（属性约简） 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，$B\subseteq C$，$0<\lambda<1$，那么 B 被称为决策信息系统的一个约简，如果它满足：

(1) $POS_B^{\lambda}(D)\subseteq POS_C^{\lambda}(D)$；

(2) $\forall b\in B, POS_{B-b}^{\lambda}(D)\neq POS_B^{\lambda}(D)$。

定义 14（属性重要度） 给定决策信息系统 $DIS=\langle U,C\cup D,V,f\rangle$，$C$ 为条件属性的集合，D 为决策属性。属性子集 $B\subseteq C$，$a\in C-B$，则 a 对 B 的重要度定义为：

$$SIG(a,B,D)=\gamma_{B\cup a}^{\lambda}(D)-\gamma_B^{\lambda}(D)\tag{12}$$

根据属性重要度，按照前向搜索的策略给出基于距离比值尺度的模糊粗糙集的属性约简算法，如算法 1 所示。

算法 1 特征选择算法

输入：DIS=⟨U,C∪D,V,f⟩,ε

输出：属性约简 E

```
1. E←∅
2. SIG_max←0
3. for each a∉C−E
4.    compute SIG(a,E,D)←γ^λ_{E∪a}(D)−γ^λ_E(D)
5.    if SIG_max<SIG(a,E,D) then
6.       a_best←a
7.       SIG_max←SIG(a,E,D)
8.    end if
9. end for
10. if  SIG_max>ε then
11.       E←E∪a_best
12.       go to 2.
13. else
14.       return E
15. end if
```

4 实验结果与分析

将基于积极样本和基于消极样本的 DRFRS 分别记为

DRFRS-F 和 DRFRS-L。为了验证所提算法的可行性及有效性，将本文提出的基于距离比值的模糊粗糙集(DRFRS-F 与 DRFRS-L)与经典粗糙集(CRS)和基于距离测度的模糊粗糙集(FRSDM)进行对比。本文实验使用分类精度作为算法的测试指标，并采用 10 折交叉验证的方法，实验结果均为交叉验证后的结果。

4.1 数据集

本文实验采用的 11 个数据集均取自于 UCI 机器学习数据库，实验数据集的信息如表 1 所列。

表 1 数据集

Table 1 Data set

Data sets	Samples	Attributes	Classes
credit. a	690	15	2
glass	214	9	7
heart. c	303	13	5
heart. statlog	270	13	2
hepatitis	155	19	2
ILPD	582	10	2
ionosphere	351	34	2
messidor	1 151	19	2
wine	178	13	3
sonar	208	60	2
wdbc	569	30	2

4.2 实验设置

为了解决经典粗糙集不能直接对数值数据进行处理的问题，此次实验采用最小描述长度原则(MDLP)的方法对数值数据进行离散化。此外，所用数据集中的数值属性值均被标准化到[0，1]区间。

实验采用特征选择算法来对原始数据进行特征选择，之后用不同的分类器对特征选择后的数据进行训练，将这一过程进行 10 折交叉验证，以确保实验的可靠性。所采用的分类器分别为支持向量机(SVM)、朴素贝叶斯(NaiveBayes)及 J48 分类器。实验中核参数 $\sigma=0.15$，特征选择算法中的控制参数 ε 被设置为 0.005，参数 k 取值为 11，实验平台为 Windows 10，使用 Java 1.8 编程，分类算法使用 Weka 3.8 提供的 API。

4.3 实验结果

为了研究距离比值尺度 λ 与特征选择属性个数的关系及分类精度随距离比值尺度 λ 的变化情况，同时验证特征约简算法的有效性，本文实验分为 3 个部分，分别为验证约简属性个数随 λ 的变化情况、分类精度随 λ 的变化情况，以及不同分类器下的算法之间的对比实验。

4.3.1 λ 与特征选择属性个数的关系

数据集 hepatitis，ionosphere，wine，sonar，wdbc 和 heart. statlog 经特征选择后的属性个数随 λ 的变化情况如图 2、图 3 所示。

由图 2 和图 3 可知，该特征选择算法能够有效减少数据集的属性个数，当 $\lambda\in[0,0.5)$ 时，数据集 ionosphere，sonar，wine 和 wdbc 经过特征选择算法处理后的约简属性个数曲线波动较小，当 $\lambda\in(0.5,1)$ 时，大部分数据集下的约简属性个数随着 λ 的增加而逐渐减少。

由图 3 可知，数据集 hepatitis 和 heart. statlog 经过特征选择算法处理后的属性个数随 λ 的波动较大，而在图 2 中这两个数据集的曲线仍然比较缓和。由此可知，DRFRS-F 相对保守，而 DRFRS-L 在部分数据集下比较敏感。

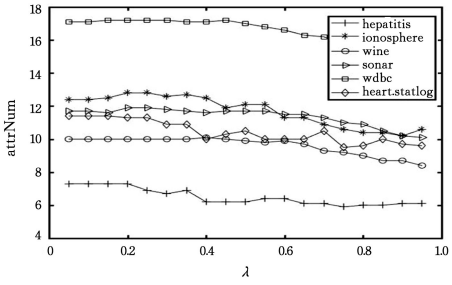


图 2 约简属性个数随 λ 的变化情况(DRFRS-F)

Fig. 2 Relationship between λ and number of attributes(DRFRS-F)

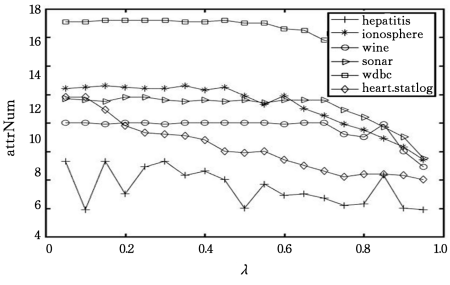
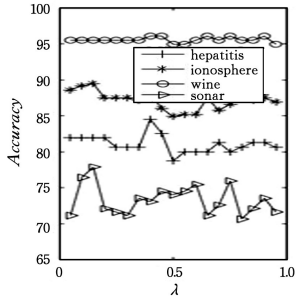


图 3 约简属性个数随 λ 的变化情况(DRFRS-L)

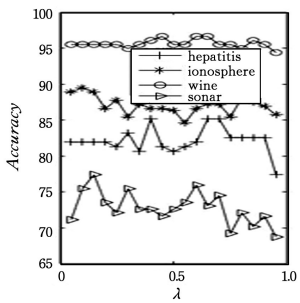
Fig. 3 Relationship between λ and number of attributes(DRFRS-L)

4.3.2 分类精度随 λ 的变化情况

数据集 hepatitis，ionosphere，wine 和 sonar 由特征选择算法过滤后，在不同分类器下的分类精度随 λ 的变化情况如图 4—图 6 所示。对比不同数据集下 λ 取值与精度的关系，发现不同数据集的最佳 λ 取值也有所不同。此外，同一数据集在不同分类器下取得最大精度时的 λ 取值也有所不同，这是由不同分类器对数据集属性的敏感程度不同造成的。



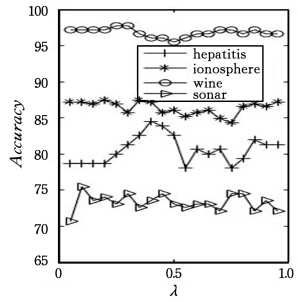
(a)DRFRS-F



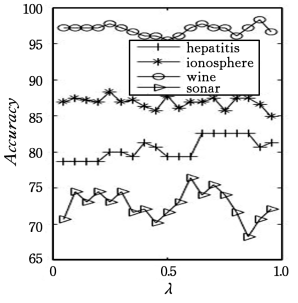
(b)DRFRS-L

图 4 分类精度 *accuracy* 随 λ 的变化情况(NaiveBayes)

Fig. 4 Relationship between λ and *accuracy* (NaiveBayes)



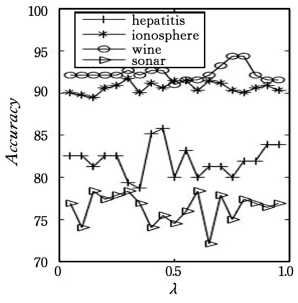
(a)DRFRS-F



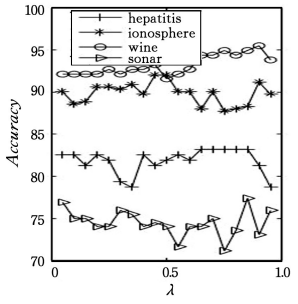
(b)DRFRS-L

图 5 分类精度 *accuracy* 随 λ 的变化情况(SVM)

Fig. 5 Relationship between λ and *accuracy* (SVM)



(a)DRFRS-F



(b)DRFRS-L

图 6 分类精度 *accuracy* 随 λ 的变化情况(J48)

Fig. 6 Relationship between λ and accuracy (J48)

4.3.3 算法对比实验结果

经过特征选择算法后的数据集，在 NaiveBayes，SVM 及 J48 下的最优参数下的分类精度如表 2－表 4 所列。其中，“－”表示在当前数据集下并不能得到有效的属性约简集合，因此无分类精度。由表 2 可知，本文提出的 DRFRS-L 特征选择算法在 credit. a，heart. c，heart. statlog 等数据集上均能获得较高的分类精度，且在所有数据集上的平均分类精度比 FRSDM 高1. 19%。

表 2 NaiveBayes 上的分类精度

Table 2 *Accuracy* on NaiveBayes

(单位：%)

Data Set	Raw	RS	FRSDM	DRFRS-F	DRFRS-L
credit. a	77.68	76.09	81.45	86.81	**87.10**
glass	48.6	—	**49.07**	48.60	48.60
heart. c	83.5	83.5	84.49	84.82	**85.15**
heart. statlog	83.7	—	84.44	84.44	**85.56**
hepatitis	84.52	**87.74**	84.52	84.52	85.16
ILPD	55.67	55.33	55.67	**57.56**	**57.56**
ionosphere	82.62	**90.6**	89.17	89.46	89.46
messidor	56.82	56.3	62.21	**63.60**	63.42
wine	96.63	96.63	**98.31**	96.07	96.63
sonar	67.79	—	72.6	**77.88**	77.40
wdbc	92.97	93.85	**95.96**	94.55	94.90
Avg.	75.50	—	77.99	78.94	**79.18**

在 SVM 下的分类精度如表 3 所列。

表 3 SVM 上的分类精度

Table 3 *Accuracy* on SVM

(单位：%)

Data Set	Raw	RS	FRSDM	DRFRS-F	DRFRS-L
credit. a	**85.51**	**85.51**	**85.51**	**85.51**	**85.51**
glass	47.66	—	47.66	**48.60**	48.60
heart. c	84.49	84.82	84.49	**85.48**	84.82
heart. statlog	83.7	—	83.7	84.07	**85.19**
hepatitis	80.65	78.06	80.65	**84.52**	82.58
ILPD	**71.31**	**71.31**	**71.31**	**71.31**	**71.31**
ionosphere	86.89	87.75	**89.46**	87.46	88.32
messidor	60.82	60.73	60.99	**61.69**	**61.69**
wine	97.75	96.63	97.75	97.75	**98.31**
sonar	70.19	—	75	75.48	**76.44**
wdbc	94.9	94.38	95.43	**95.96**	**95.96**
Avg.	78.53	—	79.27	79.80	**79.88**

由表 3 可知，在 credit. a，glass，wdbc 等数据集下，DRFRS-F 和 DRFRS-L 均获得了相同的分类精度，这是由该分类器的性质和基于距离比值尺度的样本集合共同决定的。在数据集 heart. statlog，hepatitis 和 sonar 下，DRFRS-L 相比 FRSDM 均有较大提高，且在所有数据集下的平均精度比 FRSDM 高。

在 J48 分类器下的分类精度如表 4 所列。由表 4 可知，在大多数数据集下，DRFRS-L 均能够获得较高的分类精度，且在所有数据集下的平均精度比 FRSDM 高 0. 91%。

表 4 J48 上的分类精度

Table 4 *Accuracy* on J48

(单位：%)

Data Set	Raw	RS	FRSDM	DRFRS-F	DRFRS-L
credit. a	86.09	86.23	86.09	86.52	**86.81**
glass	66.82	—	67.76	68.22	**70.09**
heart. c	77.56	76.9	79.87	78.22	**81.19**
heart. statlog	76.67	—	80.74	80.37	**81.85**
hepatitis	83.87	81.29	**85.81**	**85.81**	83.23
ILPD	65.46	65.98	65.64	**67.01**	**67.01**
ionosphere	91.45	**92.02**	91.74	91.74	**92.02**
messidor	64.38	62.9	64.38	**65.51**	65.25
wine	93.82	93.26	94.38	94.38	**95.51**
sonar	71.15	—	74.52	**78.37**	77.40
wdbc	93.15	**94.9**	94.38	94.38	94.73
Avg.	79.13	—	80.48	80.96	**81.37**

综上所述，本文提出的特征约简算法在分类器 NaiveBayes，SVM 及 J48 下均能取得较好的效果。

结束语 为了选取合适的样本以更合理地定义下近似算子，本文提出了基于距离比值尺度的模糊粗糙集，并给出了基于该模型的属性约简算法。实验结果证明了该算法的有效性，且该算法在大部分数据集下具有较高的分类精度。但仍然有些问题需要进一步探讨，如数据集对距离比值尺度的大小是参数敏感的，如何能够针对不同的数据设置合适的泛化参数范围，需要进一步的讨论。

参 考 文 献

[1] HONG R C, PAN J X, HAO S J, et al. Image quality assessment based on matching pursuit[J]. Information Sciences, 2014, 273: 196-211.

[2] HONG R C, WANG M, GAO Y, et al. Image annotation by multiple-instance learning with discriminative feature mapping and selection[J]. IEEE Transactions on Cybernetics, 2014, 44(5): 669-680.

[3] LU J J, ZHAO T Z, ZHANG Y F. Feature selection based-on genetic algorithm for image annotation[J]. Knowledge-Based Systems, 2008, 21(8): 887-891.

[4] PAWLAK Z. Rough set [J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[5] CHEN J K, LI J J, LIN Y J. Computing connected components of simple undirected graphs based on generalized rough sets[J]. Knowledge-Based Systems, 2013, 37: 80-85.

[6] CHEN H M, LI T R, LUO C, et al. A decision-theoretic rough set approach for dynamic data mining[J]. IEEE Transactions on Fuzzy Systems, 2015, 23(6): 1958-1970.

[7] CHEN J K, LIN Y, LIN G, et al. The relationship between attribute reducts in rough sets and minimal vertex covers of graphs[J]. Information Sciences, 2015, 325: 87-97.

[8] LI J H, REN Y, MEI C L, et al. A comparative study of multigranulation rough sets and concept lattices via rule acquisition [J]. Knowledge-Based Systems, 2016, 91: 152-164.

[9] DUBOIS D, PRADE H. Rough fuzzy sets and fuzzy rough sets [J]. International Journal of General Systems, 1990, 17(2/3): 191-209.

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

[11] JENSEN R, SHEN Q. Fuzzy-rough attribute reduction with application to web categorization[J]. Fuzzy Sets and Systems, 2004, 141(3): 469-485.

[12] SHEN Q, JENSEN R. Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring [J]. Pattern Recognition, 2004, 37(7): 1351-1363.

[13] HU Q H, YU D R, XIE Z X. Information-preserving hybrid data reduction based on fuzzy-rough techniques[J]. Pattern Recognition Letters, 2006, 27(5): 414-423.

[14] WANG C Z, QI Y L, HE Q. Attribute reduction using distance-based fuzzy rough sets[C]// International Conference on Machine Learning and Cybernetics. IEEE, 2015: 860-865.

[15] PAWLAK Z. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Kluwer Academic Publishers, 1992.

[16] ZHANG W X. Rough set theory and method[M]. Beijing: Science Press, 2001.

[17] YEUNG D S, CHEN D G, TSANG E C C, et al. On the generalization of fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2005, 13(3): 343-361.

[18] MORSI N N, YAKOUT M M. Axiomatics for fuzzy rough sets [J]. Fuzzy Sets and Systems, 1998, 100(1/2/3): 327-342.

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

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

CHEN Yi-ning, born in 1995, postgraduate. His main research interests include areas of rough set and so on.

CHEN Hong-mei, born in 1971, Ph. D, professor, Ph. D supervisor, is member of China Computer Federation (CCF). Her main research interests include rough set, granular computing, and intelligent information processing.