计算机科学
COMPUTER SCIENCE

基于误分代价的变精度模糊粗糙集属性约简

王子茵，李磊军，米据生，李美争，解滨

引用本文

王子茵，李磊军，米据生，李美争，解滨. 基于误分代价的变精度模糊粗糙集属性约简[J]. 计算机科学, 2022, 49(4): 161-167.

WANG Zi-yin, LI Lei-jun, MI Ju-sheng, LI Mei-zheng, XIE Bin. Attribute Reduction of Variable Precision Fuzzy Rough Set Based on Misclassification Cost[J]. Computer Science, 2022, 49(4): 161-167.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于邻域粗糙集和 Relief 的弱标记特征选择方法

Weak Label Feature Selection Method Based on Neighborhood Rough Sets and Relief

计算机科学, 2022, 49(4): 152-160. https://doi.org/10.11896/jsjkx.210300094

一种基于正域的三支近似约简

Three-way Approximate Reduction Based on Positive Region

计算机科学, 2022, 49(4): 168-173. https://doi.org/10.11896/jsjkx.210500067

带标记的不完备双论域模糊概率粗糙集中近似集动态更新方法

Label-based Approach for Dynamic Updating Approximations in Incomplete Fuzzy Probabilistic Rough Sets over Two Universes

计算机科学, 2022, 49(3): 255-262. https://doi.org/10.11896/jsjkx.201200042

基于 k-原型聚类和粗糙集的属性约简方法

Attribute Reduction Method Based on *k*-prototypes Clustering and Rough Sets

计算机科学, 2021, 48(6A): 342-348. https://doi.org/10.11896/jsjkx.201000053

基于多粒度粗糙直觉犹豫模糊集的最优粒度选择方法

Optimal Granulation Selection Method Based on Multi-granulation Rough Intuitionistic Hesitant Fuzzy Sets

计算机科学, 2021, 48(10): 98-106. https://doi.org/10.11896/jsjkx.200800074

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210500211

基于误分代价的变精度模糊粗糙集属性约简

王子茵[1,3]　李磊军[2,3,4]　米据生[2,3]　李美争[1]　解　滨[1]
1 河北师范大学计算机与网络空间安全学院　石家庄 050024
2 河北师范大学数学科学学院　石家庄 050024
3 河北省计算数学与应用重点实验室　石家庄 050024
4 河北师范大学数学博士后科研流动站　石家庄 050024
(1060034724@qq.com)

摘　要　属性约简目前是粗糙集领域的热点研究问题。文中研究了如何在保持误分类代价不增加的基础上减少冗余属性。首先定义了变精度模糊粗糙集中的最小误分类程度,然后引入了决策过程,提出了一种基于最小误分类程度的变精度模糊粗糙集模型,最后在这个模型的基础上将误分代价作为不变量,提出了一种启发式属性约简算法。将所提算法与其他算法进行比较,实验结果表明,所提算法得到的属性约简结果具有保留的属性数相对较少、误分类代价更低的优点。

关键词:粗糙集;变精度模糊粗糙集;误分类代价;属性约简
中图法分类号　TP182

Attribute Reduction of Variable Precision Fuzzy Rough Set Based on Misclassification Cost

WANG Zi-yin[1,3], LI Lei-jun[2,3,4], MI Ju-sheng[2,3], LI Mei-zheng[1] and XIE Bin[1]
1 College of Computer and Cyberspace Security, Hebei Normal University, Shijiazhuang 050024, China
2 College of Mathematical Sciences, Hebei Normal University, Shijiazhuang 050024, China
3 Hebei Key Laborotory of Computational Mathematics and Applications, Shijiazhuang 050024, China
4 Postdoctoral Research Workstation of Mathematics, Hebei Normal University, Shijiazhuang 050024, China

Abstract　Attribute reduction is a hot research issue in rough set. In this paper, how to reduce redundant attributes without increasing the misclassification cost is studied. Firstly, the minimum misclassification degree of variable precision fuzzy rough sets is defined. Then, by introducing the decision process, the variable precision fuzzy rough set model is proposed based on the minimum misclassification degree. Then, a heuristic attribute reduction algorithm is proposed by taking the misclassification cost as an invariant. We compare this algorithm with other algorithms through experiments. The results show that the attribute reduction results obtained by the proposed algorithm have the advantages of less reserved attributes and lower misclassification cost.

Keywords　Rough set, Variable precision fuzzy rough set, Misclassification cost, Attribute reduction

1　引言

粗糙集[1-2]理论是 Pawlak 于 20 世纪 80 年代提出的用于处理不确定、不精确和不完整信息的数学理论。近年来,该理论已被成功地应用于计算机科学和管理科学的许多领域,如智能数据处理[3]、数据挖掘[4]等。

由于信息系统中存在不确定性信息和噪声属性,但经典的粗糙集模型不允许任何错误,因此造成它对错误分类和扰动非常敏感。基于这一点,学者在经典粗糙集模型的基础上做了许多扩展。Yao 等[5]在贝叶斯决策理论的基础上提出了一种决策理论粗糙集(DTRS)模型,这个模型的提出促进了概率粗糙集(PRS)模型的发展[6]。Ziarko 于 1993 年[7]提出了一种基于分类误差度量的变精度粗糙集模型,这是一种处理不确定信息和提高数据对噪声抗干扰性的方法,可以降低错误决策的可能性。事实上,由于 VPRS 模型中的分类误差度量可以用 DTRS 中的条件概率表示,因此 VPRS 模型是 DTRS 模型的一种特例[6]。接下来,本文进一步探讨了模糊背景下的 DTRS 和 VPRS 模型。Zhao 等[8]研究了变精度模糊

到稿日期:2021-05-29　返修日期:2021-10-21
基金项目:国家自然科学基金(61502144,62076088);河北省自然科学基金(F2018205196,F2019205295);河北省高等学校自然科学基金(BJ2019014);河北省博士后择优资助科研项目(B2016003013);河北省三三三人才工程培养经费(A2017002112)
This work was supported by the National Natural Science Foundation of China(61502144,62076088), Natural Science Foundation of Hebei Province(F2018205196,F2019205295), Natural Science Foundation of Higher Education Institutions of Hebei Province(BJ2019014), Postdoctoral Advanced Programs of Hebei Province(B2016003013) and Training Funds for 333 Talents Project in Hebei Province(A2017002112).
通信作者:李磊军(lileijun1985@163.com)

粗糙集(VPFRS)的概念。Sun[9]等提供了一种在模糊环境下构造变精度多粒度粗糙集模型的新方法,并提出了一种基于该模型的多属性群决策方法。Yao 等[10]提出了基于模糊颗粒的变精度模糊粗糙集模型,并研究了该模型的基本性质。

此外,属性约简也引起了许多研究者的关注,并基于不同视角提出了不同属性约简的方法,如基于代数视角的属性约简[11-14]、基于信息熵的属性约简[15-17]和基于属性重要度的属性约简[18-19]。Miao 等[20]提出了一种基于互信息的相对知识约简算法。Wang 等[21]从信息和代数的角度分析了属性约简的基本原理,并进一步提出了基于条件信息熵的知识约简算法。Mi 等[22]研究了变精度粗糙集模型下的知识约简方法。此外,Liao 等[23]通过讨论阈值 α,β,γ,提出了基于决策风险最小化的属性约简。Zhang 等[18]提出了决策理论粗糙集中基于属性重要度的属性约简。Bi 等[24]提出了一种基于风险成本最小化和联合属性重要度的属性约简算法。Yu 等[25]考虑到不同属性具有不同的分类能力,提出了一种基于风险最小化的属性约简算法。此外,Liao 等[26]提出了一种代价敏感粗糙集,并给出了一个使分类成本最小化的属性约简新定义。Yang 等提出了一种使误分类代价最小化的属性约简[27]。综上所述,考虑代价敏感性的属性约简已经取得很大的成果,但很少有研究者关注区域划分带来的误分类代价。

本文基于文献[27]中的误分类代价的变精度粗糙集模型,重新构造了模糊背景下基于误分类代价的变精度粗糙集模型,将决策过程引入该模型,同时定义了模糊背景下的最小误分类程度,并在此基础上给出了误分类代价的计算方法,最后提出了一种基于误分类代价的启发式属性约简算法,并对属性约简的结果及误分代价进行了分析。

2 基础知识

定义 1[8] 三元组 $S=(U,A,V,f,d)$是一个决策信息系统,其中 $U=\{x_1,x_2,\cdots,x_n\}$是由对象组成的非空有限论域,$A=\{a_1,a_2,\cdots,a_m\}$是条件属性的有限集合,d 是决策属性,$V=\{V_a\mid a\in A\}$是属性值集,f 表示一个映射 $f:U\times A\rightarrow V_a$,如果 $V_a\subseteq[0,1]$,那么这个系统就叫做模糊决策信息系统。

定义 2[13] 设 U 是对象集,$U\times U$ 上的模糊集合 R 称为 U 上的一个模糊关系,对于 U 上的模糊关系 R:

(1)称 R 是自反的,若 $R(x,x)=1\ (\forall x\in U)$;

(2)称 R 是对称的,若 $R(x,y)=R(y,x)(\forall x,y\in U)$;

(3)称 R 是传递的,若 $R(x,z)\geqslant R(x,y)\wedge R(y,z)(\forall x,y,z\in U)$。

若 R 满足自反、对称和传递关系,则称 R 为 U 上的一个模糊等价关系,利用模糊等价关系可以对论域进行动态划分。

根据标准化数值建立模糊等价关系矩阵 $\boldsymbol{R}$,本文使用最大最小法:

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{pmatrix}=(r_{ij})_{m\times n}$$

其中,$r_{ij}=\dfrac{\sum_{k=1}^{m}\min\{a_k(x_i),a_k(x_j)\}}{\sum_{k=1}^{m}\max\{a_k(x_i),a_k(x_j)\}}$,假设决策表共有 m 个属性值,$a_k(x_i)$,$a_k(x_j)$分别表示第 i 个对象和第 j 个对象的第 k 个属性值($1\leqslant k\leqslant m$)。

定义 3[26] 给定属性集 A 和对象集 U,R_A 是 U 上由 A 诱导的模糊等价关系,$U/IND(A)=\{R_A(x_1),R_A(x_2),\cdots,R_A(x_n)\}$表示由模糊等价关系 R_A 划分的模糊等价类,也可表示为 U/A。其中,$\forall x\in U$,$R_A(x)$叫做 x 的模糊邻域,它是一个模糊集合,且满足 $R_A(x)(y)=R_A(x,y)$,$\forall y\in U$。

x 相对于模糊等价关系 R_A 的模糊邻域为:

$$R_A(x)=\sum_{y\in U}R_A(x,y)/y$$

每个对象的模糊邻域的集合构成了该对象的模糊等价类。$U/IND(D)=\{D_1,D_2,\cdots,D_n\}$表示由决策属性划分的经典等价类,也可表示为 U/D,其中 $D_j\in U/IND(D)$叫做决策类。

例 1 计算模糊等价类。

给定一个决策信息系统,如表 1 所列。

表 1 决策信息系统

Table 1 Decision information system

U	a_1	a_2	a_3	a_4	d
x_1	0.7	1.0	0.8	0.4	1.0
x_2	1.0	0.9	1.0	0.2	0
x_3	0.8	0.7	0.6	1.0	1.0
x_4	1.0	0.7	1.0	0.3	1.0
x_5	0.7	1.0	1.0	0.2	1.0
x_6	0.7	1.0	0.6	1.0	0
x_7	1.0	0.7	0.2	1.0	0
x_8	0.2	0.1	1.0	1.0	1.0

由表 1 可以看出,决策属性集 $D=\{d\}$,条件属性集 $A=\{a_1,a_2,a_3,a_4\}$,论域 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$,利用条件属性集 A 可以定义一个模糊等价关系 R:

$$R(x_i,x_j)=\frac{\sum_{k=1}^{m}\min\{a_k(x_i),a_k(x_j)\}}{\sum_{k=1}^{m}\max\{a_k(x_i),a_k(x_j)\}},\ \forall x_i,x_j\in U$$

由表 1 计算得:

$$\boldsymbol{R}=\begin{pmatrix} 1 & 0.76 & 0.67 & 0.74 & 0.87 & 0.73 & 0.53 & 0.41 \\ 0.76 & 1 & 0.59 & 0.91 & 0.88 & 0.65 & 0.54 & 0.38 \\ 0.67 & 0.59 & 1 & 0.65 & 0.58 & 0.89 & 0.82 & 0.54 \\ 0.74 & 0.91 & 0.65 & 1 & 0.79 & 0.63 & 0.59 & 0.43 \\ 0.87 & 0.88 & 0.58 & 0.79 & 1 & 0.64 & 0.45 & 0.41 \\ 0.73 & 0.65 & 0.89 & 0.63 & 0.64 & 1 & 0.78 & 0.49 \\ 0.53 & 0.54 & 0.82 & 0.59 & 0.45 & 0.78 & 1 & 0.41 \\ 0.41 & 0.38 & 0.54 & 0.43 & 0.41 & 0.49 & 0.41 & 1 \end{pmatrix}$$

决策等价类为:

$$U/D=\{D_1,D_2\}=\{\{x_1,x_3,x_4,x_5,x_8\},\{x_2,x_6,x_7\}\}$$

模糊条件等价类为:

$$\begin{aligned} U/IND(A)&=\{R_A(x_1),R_A(x_2),R_A(x_3),\cdots,R_A(x_8)\} \\ &=\{\frac{1}{x_1}+\frac{0.76}{x_2}+\frac{0.67}{x_3}+\frac{0.74}{x_4}+\frac{0.87}{x_5}+\frac{0.73}{x_6}+ \\ &\quad \frac{0.53}{x_7}+\frac{0.41}{x_8},\frac{0.76}{x_1}+\frac{1}{x_2}+\frac{0.58}{x_3}+\frac{0.91}{x_4}+ \\ &\quad \frac{0.88}{x_5}+\frac{0.65}{x_6}+\frac{0.54}{x_7}+\frac{0.38}{x_8},\frac{0.67}{x_1}+\frac{0.59}{x_2}+ \\ &\quad \frac{1}{x_3}+\frac{0.65}{x_4}+\frac{0.58}{x_5}+\frac{0.89}{x_6}+\frac{0.82}{x_7}+\frac{0.54}{x_8}, \end{aligned}$$

$$\frac{0.74}{x_1}+\frac{0.91}{x_2}+\frac{0.65}{x_3}+\frac{1}{x_4}+\frac{0.79}{x_5}+\frac{0.63}{x_6}+\frac{0.59}{x_7}+\frac{0.43}{x_8},\frac{0.87}{x_1}+\frac{0.88}{x_2}+\frac{0.58}{x_3}+\frac{0.79}{x_4}+\frac{1}{x_5}+\frac{0.64}{x_6}+\frac{0.45}{x_7}+\frac{0.41}{x_8},\frac{0.73}{x_1}+\frac{0.65}{x_2}+\frac{0.89}{x_3}+\frac{0.63}{x_4}+\frac{0.64}{x_5}+\frac{1}{x_6}+\frac{0.78}{x_7}+\frac{0.49}{x_8},\frac{0.53}{x_1}+\frac{0.54}{x_2}+\frac{0.82}{x_3}+\frac{0.59}{x_4}+\frac{0.45}{x_5}+\frac{0.78}{x_6}+\frac{1}{x_7}+\frac{0.41}{x_8},\frac{0.41}{x_1}+\frac{0.38}{x_2}+\frac{0.54}{x_3}+\frac{0.43}{x_4}+\frac{0.41}{x_5}+\frac{0.49}{x_6}+\frac{0.41}{x_7}+\frac{1}{x_8}\}$$

定义 4 在模糊决策信息系统 $S=(U,A,V,f,d)$ 中，设 X 是 U 上的一个模糊集合，$R_A(x)$ 是 x 的模糊邻域，模糊集合 $R_A(x)$ 与 X 之间的误分类程度为：

$$c(R_A(x),X)=1-\frac{|R_A(x)\cap X|}{|R_A(x)|}$$

其中，$|X|=\sum_{x\in U}X(x)$，误分类程度指模糊集合 $R_A(x)$ 关于模糊集合 X 的错误分类率。

定义 5[8] 在模糊决策信息系统 $S=(U,A,V,f,d)$ 中，R_A 是 U 上由 A 诱导的模糊等价关系，设 X 是论域 U 上的一个模糊集合，令 $0<\alpha\leqslant 0.5$，定义变精度模糊粗糙集的上、下近似为：

$$\underline{R}^{\alpha}(X)=\{x:c(R_A(x),X)\leqslant\alpha\}$$

$$\overline{R}^{\alpha}(X)=\{x:c(R_A(x),X)<1-\alpha\}$$

X 的正域、负域、边界域分别为：

$$POS^{\alpha}(X)=\underline{R}^{\alpha}(X)$$

$$BND^{\alpha}(X)=\overline{R}^{\alpha}(X)-\underline{R}^{\alpha}(X)=\{x:\alpha<c(R_A(x),X)<1-\alpha\}$$

$$NEG^{\alpha}(X)=U-\overline{R}^{\alpha}(X)=\{x:c(R_A(x),X)\geqslant 1-\alpha\}$$

3 变精度模糊粗糙集中基于误分类代价的属性约简

3.1 变精度模糊粗糙集中的误分类代价

本节定义了变精度模糊粗糙集中的最小误分类程度，然后引入决策过程，提出了一种基于最小误分类程度的变精度模糊粗糙集模型，进而在这个模型的基础上提出了一种启发式属性约简算法。

定义 6 $R_A(x_j)$ 相对于 U/D 的最小误分类程度定义为：

$$c(R_A(x_j),U/D)_{\min}=\min_{i=1}^{n}c(R_A(x_j),D_i)$$

其中，$c(R_A(x_j),D_i)=1-\frac{|R_A(x_j)\cap D_i|}{|R_A(x_j)|}$，$c(R_A(x),U/D)_{\min}=\{c(R_A(x_1),U/D)_{\min},c(R_A(x_2),U/D)_{\min},\cdots,c(R_A(x_m),U/D)_{\min}\}$。下文用 $c_{\min}$ 来表示最小误分类程度。

例 2 根据定义 4 和例 1 求得的模糊条件等价类，可以得到任意模糊邻域相对于每个决策类的误分类程度。

$$c(R_A(x_1),D_1)=1-\frac{3.68}{5.79}=0.36$$

$$c(R_A(x_1),D_2)=0.64$$

$$c(R_A(x_2),D_1)=0.38,c(R_A(x_2),D_2)=0.62$$

$$c(R_A(x_3),D_1)=0.42,c(R_A(x_3),D_2)=0.58$$

$$c(R_A(x_4),D_1)=0.37,c(R_A(x_4),D_2)=0.63$$

$$c(R_A(x_5),D_1)=0.35,c(R_A(x_5),D_2)=0.65$$

$$c(R_A(x_6),D_1)=0.41,c(R_A(x_6),D_2)=0.59$$

$$c(R_A(x_7),D_1)=0.45,c(R_A(x_7),D_2)=0.55$$

$$c(R_A(x_8),D_1)=0.32,c(R_A(x_8),D_2)=0.68$$

根据定义 6 可以得到每个对象的模糊邻域相对于 U/D 的最小误分类程度：

$$c(R_A(x_1),U/D)_{\min}=0.36$$

$$c(R_A(x_2),U/D)_{\min}=0.38$$

$$c(R_A(x_3),U/D)_{\min}=0.42$$

$$c(R_A(x_4),U/D)_{\min}=0.32$$

$$c(R_A(x),U/D)_{\min}=\{0.36,0.38,0.42,0.32\}$$

定理 1 对于 x_j 的模糊邻域 $R_A(x)$ 和任意决策类 $D_i\in U/D$，满足：

(1)若 $R_A(x_j)\subseteq D_i$，则 $c(R_A(x_j),D_i)=0$；

(2)若 $\exists x\in U$ 使 $R_A(x_j)(x)>D_i(x)$，则 $0<c(R_A(x_j),D_i)<1$；

(3)$0\leqslant c(R_A(x_j),U/D)_{\min}<1$。

证明：(1)若 $R_A(x_j)\subseteq D_i$，则 $R_A(x_j)(x)\leqslant D_i(x)$，那么 $|R_A(x_j)\cap D_i|=|R_A(x_j)|$，因此 $c(R_A(x_j),D_i)=1-\frac{|R_A(x_j)\cap D_i|}{|R_A(x_j)|}=0$。

(2)由定义 6 可知，$0\leqslant c(R_A(x_j),D_i)<1$，若 $\exists x\in U$ 使 $R_A(x_j)(x)>D_i(x)$，则 $|R_A(x_j)\cap D_i|\neq|R_A(x_j)|$，且 $c(R_A(x_j),D_i)\neq 0$，因此 $0<c(R_A(x_j),D_i)<1$。

(3)$0\leqslant c(R_A(x_j),D_i)<1$，故 $0\leqslant c(R_A(x_j),U/D)_{\min}<1$。

下文设 D_i 是目标概念，条件属性子集 $B\subseteq A$，给定 3 个决策动作 $\tau=\{a_P,a_B,a_N\}$。

a_P：$x\in POS_B^{\alpha}(D_i)$ 表示 x 分到 D_i 的正域中；

a_B：$x\in BND_B^{\alpha}(D_i)$ 表示 x 分到 D_i 的负域中；

a_N：$x\in NEG_B^{\alpha}(D_i)$ 表示 x 分到 D_i 的边界域中。

基于 3 个决策动作给出了 6 个损失函数，λ_{PP}，λ_{BP}，λ_{NP} 分别表示在目标概念 D_i 中的对象采取行动 a_P，a_B，a_N 的损失函数；λ_{PN}，λ_{BN}，λ_{NN} 分别表示不在目标概念 D_i 中的对象采取行动 a_P，a_B，a_N 的损失函数，如表 2 所列。

表 2 损失函数

Table 2 Loss function

	D_i	$\neg D_i$
a_P	λ_{PP}	λ_{PN}
a_B	λ_{NP}	λ_{NB}
a_N	λ_{NP}	λ_{NN}

定义 7 采取不同决策动作(将对象分到正域、负域、边界域)的误分类代价如下。

正域的代价为：

$$COST_{POS}=\sum_{c_{\min}\leqslant\alpha}[\lambda_{PP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{PN}\cdot c(R_A(x_j),U/D)_{\min}]$$

边界域的代价为：

$$COST_{BND}=\sum_{\alpha\leqslant c_{\min}\leqslant 1-\alpha}[\lambda_{BP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{BN}\cdot c(R_A(x_j),U/D)_{\min}]$$

负域的代价为：

$$COST_{NEG}=\sum_{c_{\min}\geqslant 1-\alpha}[\lambda_{NP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{NN}\cdot c(R_A(x_j),U/D)_{\min}]$$

总体的代价为：

$$COST_B^{\alpha}=COST_{POS}+COST_{BND}+COST_{NEG}$$

下面给出单一决策类的误分类代价。

正域的代价为：

$$COST_{POS}(D_i)=\sum_{c\leqslant\alpha}[\lambda_{PP}\cdot(1-c(R_A(x_j),D_i))+\lambda_{PN}\cdot c(R_A(x_j),D_i)]$$

边界域的代价为：

$$COST_{BND}(D_i)=\sum_{\alpha\leqslant c\leqslant 1-\alpha}[\lambda_{BP}\cdot(1-c(R_A(x_j),D_i))+\lambda_{BN}\cdot c(R_A(x_j),D_i)]$$

负域的代价为：

$$COST_{NEG}(D_i)=\sum_{c\geqslant 1-\alpha}[\lambda_{NP}\cdot(1-c(R_A(x_j),D_i))+\lambda_{NN}\cdot c(R_A(x_j),D_i)]$$

基于单一决策类 D_i 的总体代价为：

$$COST_B^{\alpha}(D_i)=COST_{POS}(D_i)+COST_{BND}(D_i)+COST_{NEG}(D_i)$$

误分代价最小的决策规则为：

a_P：$x\in POS_B^{\alpha}(D_i)$，若 $COST_{POS}\leqslant COST_{BND}$ 且 $COST_{POS}\leqslant COST_{NEG}$

a_B：$x\in BND_B^{\alpha}(D_i)$，若 $COST_{BND}\leqslant COST_{POS}$ 且 $COST_{BND}\leqslant COST_{NEG}$

a_N：$x\in NEG_B^{\alpha}(D_i)$，若 $COST_{NEG}\leqslant COST_{POS}$ 且 $COST_{NEG}\leqslant COST_{BND}$

决策规则也可以写成：

若 $c(R_A(x_j),D_i)\leqslant\alpha$，$c(R_A(x_j),D_i)\leqslant 1-\gamma$，则 $x\in POS_B^{\alpha}(D_i)$；若 $c(R_A(x_j),D_i)\geqslant 1-\alpha$，且 $c(R_A(x_j),D_i)\geqslant 1-\gamma$，则 $x\in NEG_B^{\alpha}(D_i)$；若 $\alpha\leqslant c(R_A(x_j),D_i)\leqslant 1-\alpha$，则 $x\in BND_B^{\alpha}(D_i)$。

其中，

$$\alpha=\frac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$$

$$1-\alpha=\beta=\frac{\lambda_{BN}-\lambda_{NN}}{(\lambda_{NP}-\lambda_{BP})+(\lambda_{BN}-\lambda_{NN})}$$

$$\gamma=\frac{\lambda_{PN}-\lambda_{NN}}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{PP})}=\frac{1}{2}$$

通常，$\lambda_{PP}=\lambda_{NN}=0$。也就是说，若一个对象 x 属于 D_i，那么 x 被分到 D_i 的正域的代价为 0；若对象 x 不属于 D_i，那么 x 被分到 D_i 的负域的代价为 0。

对损失函数加强约束条件，使得 $\begin{cases}\lambda_{NP}=\lambda_{PN}\\ \lambda_{BP}=\lambda_{BN}\\ \lambda_{PP}=\lambda_{NN}\end{cases}$，计算得 $\alpha=\frac{\lambda_{BP}}{\lambda_{PN}}$，$1-\alpha=\frac{\lambda_{NP}-\lambda_{BP}}{\lambda_{NP}}$。

从这个过程可以看出，每个对象根据其对应的最小误分类程度的大小计算得到此对象分到正域、负域和边界域的代价，再通过比较 3 个代价的大小采取相应的决策行动。

定义 8（变精度模糊粗糙集的误分类代价） 在模糊决策系统中，给定属性子集 $B\subseteq A$，设 $0<\alpha\leqslant 0.5$，变精度模糊粗糙集的误分类代价定义为：

$$COST_B^{\alpha}=\sum_{POS_B^{\alpha}(X)}[\lambda_{PP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{PN}\cdot c(R_A(x_j),U/D)_{\min}]+\sum_{BND_B^{\alpha}(X)}[\lambda_{BP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{BN}\cdot c(R_A(x_j),U/D)_{\min}]+\sum_{NEG_B^{\alpha}(X)}[\lambda_{NP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{NN}\cdot c(R_A(x_j),U/D)_{\min}]$$

由于 $\lambda_{PP}=\lambda_{NN}=0$，误分类代价可以写为：

$$COST_B^{\alpha}=\sum_{POS_B^{\alpha}(X)}\lambda_{PN}\cdot c(R_A(x_j),U/D)_{\min}+\sum_{BND_B^{\alpha}(X)}[\lambda_{BP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{BN}\cdot c(R_A(x_j),U/D)_{\min}]+\sum_{NEG_B^{\alpha}(X)}\lambda_{NP}\cdot(1-c(R_A(x_j),U/D)_{\min})$$

定理 2 在模糊决策信息系统 $S=(U,A,V,f,d)$ 中，U 是非空有限论域，决策类 $U/D=\{D_1,D_2,\cdots,D_n\}$，对任意属性子集 $B\subseteq A$，有 $COST_B^{\alpha}\leqslant COST_B^{\alpha}(D_i)$。

证明：由 $D_i\in U/D$ 可知：

$c(R_A(x_j),U/D)_{\min}\leqslant c(R_A(x_j),D_i)$

对于正域来说：

$$COST_{POS}=\sum_{c_{\min}\leqslant\alpha}\lambda_{PN}\cdot c(R_A(x_j),U/D)_{\min}$$

$$COST_{POS}(D_i)=\sum_{c\leqslant\alpha}\lambda_{PN}\cdot c(R_A(x_j),D_i)$$

由关系 $0<c_{\min}\leqslant c<\alpha$ 可知 $COST_{POS}\leqslant COST_{POS}(D_i)$。

对于边界域来说：

$$COST_{BND}=\sum_{\alpha\leqslant c_{\min}\leqslant 1-\alpha}[\lambda_{BP}\cdot(1-c(R_A(x_j),U/D)_{\min})+\lambda_{BN}\cdot c(R_A(x_j),U/D)_{\min}]$$

$$COST_{BND}(D_i)=\sum_{\alpha\leqslant c\leqslant 1-\alpha}[\lambda_{BP}\cdot(1-c(R_A(x_j),D_i))+\lambda_{BN}\cdot c(R_A(x_j),D_i)]$$

由关系 $\alpha<c_{\min}\leqslant c<1-\alpha$ 可知 $COST_{BND}\leqslant COST_{BND}(D_i)$。

对于负域来说：

$$COST_{NEG}=\sum_{c_{\min}\geqslant 1-\alpha}\lambda_{NP}\cdot(1-c(R_A(x_j),U/D)_{\min})$$

$$COST_{BND}(D_i)=\sum_{\alpha\leqslant c\leqslant 1-\alpha}\lambda_{BP}\cdot(1-c(R_A(x_j),D_i))$$

由关系 $1-\alpha<1-c\leqslant 1-c_{\min}$ 可知 $COST_{NEG}\leqslant COST_{NEG}(D_i)$。

综上，$COST_B^{\alpha}\leqslant COST_B^{\alpha}(D_i)$。

定义 9[27] 在模糊决策信息系统 $S=(U,A,V,f,d)$ 中，$U=\{x_1,x_2,\cdots,x_n\}$ 是对象集，决策类 $U/D=\{D_1,D_2,\cdots,D_n\}$，属性子集 $B\subseteq A$，$\forall a_i(a_i\in A,1\leqslant i\leqslant|A|)$ 的属性重要度定义为：

$$I_{COST}^{\alpha}(a_i)=COST_{B-\{a_i\}}^{\alpha}-COST_B^{\alpha}$$

定义 10 在模糊决策信息系统 $S=(U,A,V,f,d)$ 中，对任意属性子集 $B\subseteq A$，称 B 是 A 的一个属性约简，当且仅当满足以下两个条件：

(1) $COST_B^{\alpha}\leqslant COST_A^{\alpha}$；

(2) $\forall B'\subset B$，$COST_{B'}^{\alpha}>COST_B^{\alpha}$。

此方法保持约简后误分类代价不增加，并且对 $\forall a_i(a_i\in B)$，如果 $COST_{B-\{a_i\}}^{\alpha}>COST_B^{\alpha}$，那么属性 a_i 不被约简。

3.2 属性约简

基于 3.1 节提出的模型，给出了一种启发式属性约简算法，如算法 1 所示。

算法 1 变精度模糊粗糙集模型中基于误分类代价的属性约简算法

输入：决策信息系统 S=(U,A,V,f,d)

输出：约简 red

1. 初始化使 red=A，B=∅，随机生成 λ_{BP}，λ_{PN}，满足 $\alpha=\frac{\lambda_{BP}}{\lambda_{PN}}$，且 $\alpha\in(0,0.5]$。
2. 计算 c_{min}（每个等价类的最小误分度）。
3. 计算变精度模糊粗糙集的误分类代价 $COST^{\alpha}_{red}$，然后对 $\forall a_i\in red$ $(i=1,2,\cdots,|red|)$，计算 a_i 的属性重要度，$I^{\alpha}_{COST}(a_i)=COST^{\alpha}_{red-\{a_i\}}-COST^{\alpha}_{red}$。
4. 将所有属性按属性重要度从大到小排序，组成一个新的属性集 B。
   4.1. $\forall a_i\in B(i=1,2,\cdots,|B|)$，计算 red 中去掉 a_i 的误分类代价 $COST^{\alpha}_{red-\{a_i\}}$。
   4.2. 若 $COST^{\alpha}_{red-\{a_i\}}>COST^{\alpha}_{red}$，则 red=red，否则 red=red−$\{a_i\}$，循环计算直到 i>|C|，结束。
5. 返回约简集 red。

例 3 属性约简实例。

根据表 1 给出的模糊决策表，论域 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$，条件属性集 $A=\{a_1,a_2,a_3,a_4\}$。

设 $\lambda_{PN}=\lambda_{NP}=10$，$\lambda_{BP}=\lambda_{BN}=4$，$\alpha=\frac{\lambda_{BP}}{\lambda_{PN}}=0.4$，

$U/D=\{D_1,D_2\}=\{\{x_1,x_3,x_4,x_5,x_8\},\{x_2,x_6,x_7\}\}$

步骤 1 初始化使 $B=A$，根据定义 2 和定义 3 可得：

$$\boldsymbol{R}_B=\begin{pmatrix}1&0.76&0.67&0.74&0.87&0.73&0.53&0.41\\0.76&1&0.59&0.91&0.88&0.65&0.54&0.38\\0.67&0.59&1&0.65&0.58&0.89&0.82&0.54\\0.74&0.91&0.65&1&0.79&0.63&0.59&0.43\\0.87&0.88&0.58&0.79&1&0.64&0.45&0.41\\0.73&0.65&0.89&0.63&0.64&1&0.78&0.49\\0.53&0.54&0.82&0.59&0.45&0.78&1&0.41\\0.41&0.38&0.54&0.43&0.41&0.49&0.41&1\end{pmatrix}$$

$$\boldsymbol{R}_{B-\{a_1\}}=\begin{pmatrix}1&0.79&0.60&0.75&0.83&0.71&0.46&0.43\\0.79&1&0.52&0.86&0.95&0.57&0.38&0.45\\0.61&0.52&1&0.59&0.50&0.88&0.83&0.63\\0.75&0.86&0.59&1&0.83&0.53&0.44&0.52\\0.83&0.95&0.50&0.83&1&0.60&0.37&0.43\\0.71&0.57&0.88&0.53&0.60&1&0.73&0.57\\0.46&0.38&0.83&0.44&0.37&0.73&1&0.48\\0.43&0.45&0.62&0.52&0.43&0.57&0.48&1\end{pmatrix}$$

$$\boldsymbol{R}_{B-\{a_2\}}=\begin{pmatrix}1&0.71&0.65&0.75&0.81&0.63&0.46&0.52\\0.71&1&0.53&0.96&0.86&0.57&0.47&0.47\\0.65&0.53&1&0.57&0.54&0.96&0.77&0.64\\0.75&0.96&0.57&1&0.83&0.60&0.50&0.50\\0.81&0.86&0.54&0.83&1&0.52&0.37&0.52\\0.63&0.57&0.96&0.60&0.52&1&0.81&0.62\\0.46&0.47&0.77&0.50&0.37&0.81&1&0.47\\0.52&0.47&0.64&0.50&0.52&0.62&0.47&1\end{pmatrix}$$

$$\boldsymbol{R}_{B-\{a_3\}}=\begin{pmatrix}1&0.75&0.64&0.71&0.90&0.72&0.60&0.26\\0.75&1&0.59&0.86&0.82&0.67&0.66&0.17\\0.64&0.59&1&0.67&0.57&0.86&0.93&0.52\\0.71&0.86&0.67&1&0.70&0.63&0.74&0.22\\0.90&0.82&0.57&0.70&1&0.66&0.53&0.19\\0.72&0.67&0.86&0.63&0.66&1&0.87&0.45\\0.60&0.66&0.93&0.74&0.53&0.87&1&0.48\\0.26&0.17&0.52&0.22&0.19&0.45&0.48&1\end{pmatrix}$$

$$\boldsymbol{R}_{B-\{a_4\}}=\begin{pmatrix}1&0.80&0.77&0.73&0.93&0.85&0.57&0.41\\0.80&1&0.72&0.93&0.87&0.80&0.66&0.45\\0.77&0.72&1&0.78&0.71&0.84&0.74&0.36\\0.73&0.93&0.78&1&0.80&0.73&0.77&0.48\\0.93&0.87&0.71&0.80&1&0.79&0.53&0.48\\0.85&0.80&0.84&0.73&0.79&1&0.69&0.31\\0.57&0.66&0.74&0.70&0.53&0.69&1&0.19\\0.41&0.45&0.36&0.48&0.48&0.31&0.16&1\end{pmatrix}$$

步骤 2 根据定义 7 可以分别得到不同属性子集求得的模糊邻域的最小误分度。

$$c(R_B(x),U/D)_{min}=\{0.35,0.38,0.40,0.37,0.35,0.42,0.46,0.31\}$$

$$c(R_{B-\{a_1\}}(x),U/D)_{min}=\{0.35,0.35,0.40,0.33,0.35,0.41,0.45,0.33\}$$

$$c(R_{B-\{a_2\}}(x),U/D)_{min}=\{0.33,0.37,0.40,0.36,0.32,0.41,0.47,0.33\}$$

$$c(R_{B-\{a_3\}}(x),U/D)_{min}=\{0.37,0.42,0.41,0.40,0.37,0.43,0.43,0.34\}$$

$$c(R_{B-\{a_4\}}(x),U/D)_{min}=\{0.37,0.39,0.39,0.38,0.36,0.41,0.46,0.26\}$$

步骤 3 根据定义 8 可以得到原始的误分类代价和删除属性后的误分类代价。

$$COST^{\alpha}_B=10\times(0.35+0.38+0.37+0.35+0.31)+4\times(0.4+0.42+0.46)+4\times((1-0.4)+(1-0.42)+(1-0.46))=29.6$$

$$COST^{\alpha}_{B-\{a_1\}}=29.1,COST^{\alpha}_{B-\{a_2\}}=29.1$$

$$COST^{\alpha}_{B-\{a_3\}}=30.8,COST^{\alpha}_{B-\{a_4\}}=29.6$$

$$COST^{\alpha}_{B-\{a_3\}}>COST^{\alpha}_B,COST^{\alpha}_{B-\{a_4\}}>COST^{\alpha}_B$$

$$COST^{\alpha}_{B-\{a_1\}}<COST^{\alpha}_B,COST^{\alpha}_{B-\{a_2\}}<COST^{\alpha}_B$$

因此，条件属性集 A 相对于决策属性集 D 的可约属性是 a_1,a_2，得到约简子集 $B=\{a_3,a_4\}$。

4 实验分析

本文的所有实验数据均来自 UCI 数据库[28]，具体描述如表 3 所列。此外，实验所用数据集的属性值均用离差标准化方法归一化到[0,1]区间。

表 3 数据集描述

Table 3 Dataset description

数据集	样本个数	属性个数	数据类别数目
wine	178	13	3
hepatitis	155	19	2
heart. statlog	270	13	2

在 $\alpha=\frac{\lambda_{BP}}{\lambda_{PN}}$ 且 $\alpha\in(0,0.5]$ 的条件下，随机生成 λ_{BP}，λ_{PN}，分别基于最小误分类代价、单个决策类的最小误分类代价和正域进行属性约简，为了方便表示，将 3 种属性约简方法进行简化表达，如表 4 所列。

表 4 对比实验算法

Table 4 Comparative experiments algorithm

简化表示	算法
Alg1	基于最小误分类代价的属性约简
Alg2	基于单个决策类最小误分类代价的属性约简
Alg3	基于正域的属性约简

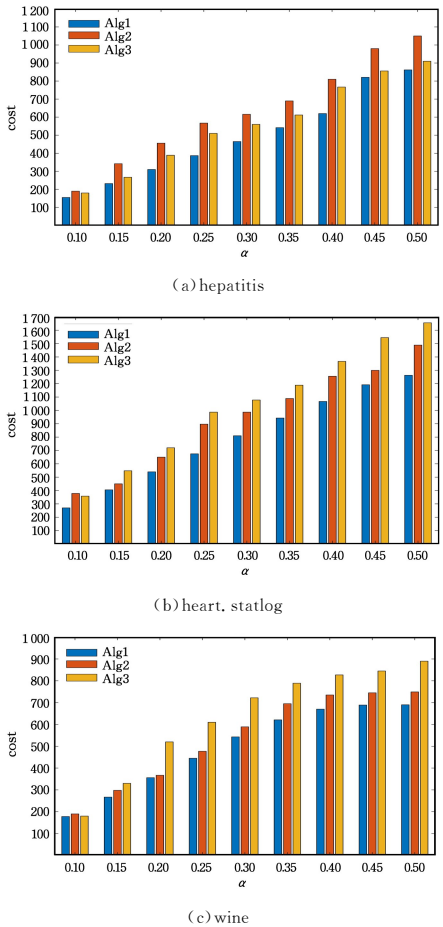
3 个数据集通过不同的方法进行属性约简后保留的属性个数如表 5 所列。

表 5 约简后属性个数对比

Table 5 Comparison of attribute number after reduction

数据集	样本数	属性数	约简后保留的属性数		
			Alg1	Alg2	Alg3
wine	178	13	5	7	3
hepatitis	155	19	7	9	6
heart. statlog	270	13	4	6	4

在决策过程中正确分类的代价为 0，即 $\lambda_{PP}=\lambda_{NN}=0$，在接下来的实验中假设 $\lambda_{PN}=\lambda_{NP}=10$，阈值 α 的大小随 λ_{BP} 和 λ_{BN} 的变化而改变。下面通过实验得到 3 个数据集的误分类代价对比图，如图 1 所示。



(a) hepatitis

(b) heart. statlog

(c) wine

图 1 误分类代价对比

Fig. 1 Comparison of misclassification costs

从表 5 和图 1 可以看出，3 个数据集通过算法 1 和算法 3 进行属性约简保留的属性个数相差不多，均比算法 2 保留的属性个数少。在 wine 数据集中，用算法 1 进行属性约简的误分类代价最小，其次是算法 2，算法 3 的误分类代价最大；在数据集 hepatitis 中，通过算法 1 进行属性约简得到的误分类代价最小，其次是算法 3，算法 2 得到的误分类代价最大；在数据集 heart. statlog 中，通过算法 1 进行属性约简得到的误分类代价最小，其次是算法 2，算法 3 得到的误分类代价最大。对于算法 1 和算法 2 来说，这是由于通过定理 2 证得对任意属性子集均有总体的误分类代价小于单个决策类的误分类代价；对于算法 1 和算法 3 来说，算法 1 将误分类代价作为不变量进行属性约简，而算法 2 在属性约简过程中没有考虑误分类代价这个指标。总的来说，与其他算法相比，算法 1 得到的属性约简结果具有保留的属性个数相对较少，误分类代价更低的优点，并且当阈值 α 增大时，容错率增加，该算法的优势更加明显。

结束语 属性约简是在保持分类能力不变的前提下简化决策信息系统的一种方法，在实际应用中利用属性约简可以提高决策的精度、速度。本文定义了最小误分类程度，在此基础上给出了误分类代价的计算方法，提出了基于误分类代价的变精度模糊粗糙集模型，并基于该模型提出了一种属性约简算法。该算法考虑了误分类代价，扩展了属性约简的应用。实验结果表明，该算法得到的属性集较小，误分类代价较低，但是该算法得到的属性约简结果对应的分类精度没有得到进一步的验证。因此，在未来的工作中，我们会从提升泛化性能的视角对该算法进行详细的讨论。

参 考 文 献

[1] PAWLAK Z. Rough Sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[2] PAWLAK Z. Rough Sets: Theoretical Aspects of Reasoning About Data[J]. Kluwer Academic Publishers, London, 1991, 97(5): 741-742.

[3] PAWLAK Z, SKOWRON A. Rudiments of Rough Sets[J]. Information Sciences, 2007, 177(1): 3-27.

[4] ZHOU J U, GUO Y. Data mining method based on rough set and fuzzy neural network[J]. Journal of Intelligent & Fuzzy Systems, 2020, 38(4): 3717-3725.

[5] YAO Y Y, WONG S K M. A decision theoretic framework for approximating concepts[J]. International Journal of Man-Machine Studies, 1992, 37(6): 793-809.

[6] YAO Y Y, GRECO S, SLOWINSKI R. Probabilistic rough sets [J]. Handbook of Computational Intelligence, 2015(2): 387-411.

[7] ZIARKO W. Variable precision rough set model[J]. Journal of Computer and System Sciences, 1993, 46: 39-59.

[8] ZHAO S Y, TSANG E C C, CHEN D G. The model of fuzzy variable precision rough sets[J]. IEEE Transactions on Fuzzy Systems, 2009, 17: 451-467.

[9] SUN B Z, MA W M, CHEN X T. Variable precision multigranulation rough fuzzy set approach to multiple attribute group decision-making based on λ-similarity relation[J]. Computers and

Industrial Engineering,2019,127:326-343.

[10] YAO Y Y,MI J S,LI Z J. A novel variable precision-fuzzy rough set model based on fuzzy granules[J]. Fuzzy Sets and Systems,2014,236:58-72.

[11] KONECNY J. On attribute reduction in concept lattices:Methods based on discernibility matrix are outperformed by basic clarifification and reduction[J]. Information Sciences,2017,415-416:199-212.

[12] WANG X,PENG Z H,LI J Y,et al. Method of Concept Reduction Based on Concept Discernibility Matrix[J]. Computer Science,2021,48(1):125-130.

[13] CHEN D G,XU W H,LI J H,et al. Basic course of granular computing[M]. Beijing:Science Press,2019.

[14] ZHANG W X,LIANG Y,WU W Z. Information system and knowledge discovery[M]. Beijing:Science Press,2014.

[15] JIANG F,WANG S S,DU J W,et al. Attribute reduction based on approximation decision entropy[J]. Control and Decision,2015,30(1):65-70.

[16] SUN L,XU J C,TIAN Y. Feature selection using roughentropy-based uncertainty measures in incomplete decision systems[J]. Knowledge-Based Systems,2012,36(6):206-216.

[17] WANG G Y. Rough set theory and knowledge acquisition[M]. Xi'an:Xi'an Jiaotong University Press,2001.

[18] ZHANG Q H,HU D R. Attribute reduction of rough set of risk decision based on attribute importance[J]. Control and Decision,2016,31(7):1199-1205.

[19] YAKOWITZ D S,LANEL J,SZIDAROVSZKY F. Multi-attribute decision making:dominance with respect to an importance order of the attributes[J],Climate Dynamics,2013,40(1/2):21-38.

[20] MIAO D,HU G R. A heuristic algorithm for reduction of knowledge[J],Journal of Computer Research and Development,1999,36(6):681-684.

[21] WANG G Y,YU H,YANG D C. Decision table reduction based on conditional information entropy[J]. Chinese Journal of Computers,2002,25(7):759-766.

[22] MI J S,WU W Z,ZHANG W X. Approaches to knowledge reduction based on variable precision rough set model[J]. Information Sciences,2004,159(3/4):255-272.

[23] JIA X Y,LIAO W H,TANG Z M,et al. Minimumcost attribute reduction in decision-theoretic rough set models[J]. Information Sciences,2013,219(1):151-167.

[24] BI Z Q,XU F F,LEI J S,et al. Attribute reduction in decision-theoretic rough set model based on minimum decision cost[J]. Concurrency and Computation Practice and Experience,2016,28(15):4125-4143.

[25] YU H,YAO Y,ZHAO J. An attribute reduction algorithm based on risk minimization[J]. Journal of Nanjing University 2013,49(49):210-216.

[26] LIAO S J,ZHU Q X,MIN F. Cost-sensitive attribute reduction in decision-theoretic rough set models[J]. Mathematical Problems in Engineering,2014,2014(1):1-9.

[27] YANG J J,ZHANG Q H,XIE Q. Attributereduction basedon misclassification cost in variable precision rough set model[J]. Journalof Intelligent & Fuzzy Systems,2019,37(4):5129-5142.

[28] BLAKE C,KEOGH E,MERZ C J. UCI Repository of Machine Learning Databases[OL]. http://archive.ics.uci.edu/ml.

WANG Zi-yin, born in 1997, master. Her main research interests include granular computing and rough set.

LI Lei-jun, born in 1985, Ph.D, associate professor, is a member of China Computer Federation and CAAI-CGCKD. His main research interests include granular computing and ensemble learning.

(责任编辑:喻藜)