专题讲座之模糊粗糙集理论

安 爽

2020.01.07

目　录

粗糙集理论是一种刻划不完整性和不确定性的数学工具， 能有效地分析不精确、不一致和不完整等各种不完备的信息。

还可以对数据进行分析和推理，从中发现隐含的知识， 揭示潜在的规律。

一、粗糙集理论概述

1. 粗糙集理论的产生

1982年，Z.Pawlak发表经典论文“Rough Sets”，标志着该理论正式诞生。

1991年，Z.Pawlak的第一本关于粗糙集理论的专著《Rough sets：theoretical aspects of reasoning about data》。

2. 粗糙集理论的特点

- 粗糙集理论是一种数据分析工具。它用确定的方法表达和处理不确定的信息和数据；能在保留关键信息的前提下对数据进行简化并求得知识的最小表达等。
- 粗糙集理论不需要先验知识。它完全由数据驱动，不需要像贝叶斯方法等统计方法的先验概率，仅利用数据本身提供的信息进行知识分类。
- 粗糙集理论是一种软计算方法。传统的硬计算使用精确、固定不变的算法来表达和解决问题，软计算是利用所允许的不精确、不确定和部分真实性以得到易于处理的、鲁棒的、成本较低的解决方案。

3. 粗糙集理论在数据挖掘中的应用

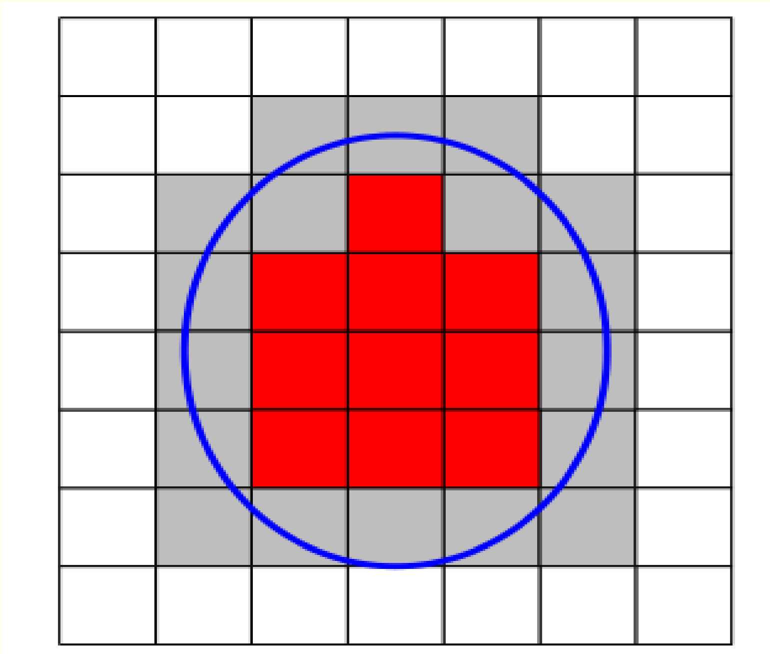
- 在数据预处理过程中，粗糙集理论可以用于对特征更准确的提取。
- 在数据准备过程中，利用粗糙集理论的数据约简特性，对数据集进行降维操作。
- 在数据挖掘阶段，可将粗糙集理论用于分类规则的发现。
- 在解释与评估过程中，粗糙集理论可用于对所得到的结果进行统计评估。

二、经典粗糙集模型

- Pawlak粗糙集
- 邻域粗糙集
- 模糊粗糙集

1. Pawlak粗糙集

Pawlak粗糙集定义 设K=（U，R）是一个知识库，对于时，称R是可定义的，又称X为R精确集；否则称R是不可定义的，也称X为R粗糙集。



U	颜色	形状	大小
1	红色	圆形	小
2	蓝色	方形	大
3	红色	三角形	小
4	蓝色	三角形	小
5	黄色	圆形	小
6	黄色	方形	小
7	红色	三角形	大
8	蓝色	三角形	大

U/A={{1},{2},{3},{4},{5},{6},{7},{8}}

定义 设*K*=（*U*，*R*）是一个知识库，对于一个集合*X*⊆*U*，则集合*X*的*R*下、上近似（集）定义为

$$\underline{R}(X)=\cup\{[x]_R \mid [x]_R \subseteq U \wedge [x]_R \subseteq X\}$$

$$\overline{R}(X)=\cup\{[x]_R \mid [x]_R \subseteq U \wedge [x]_R \cap X \neq 0\}$$

或定义如下

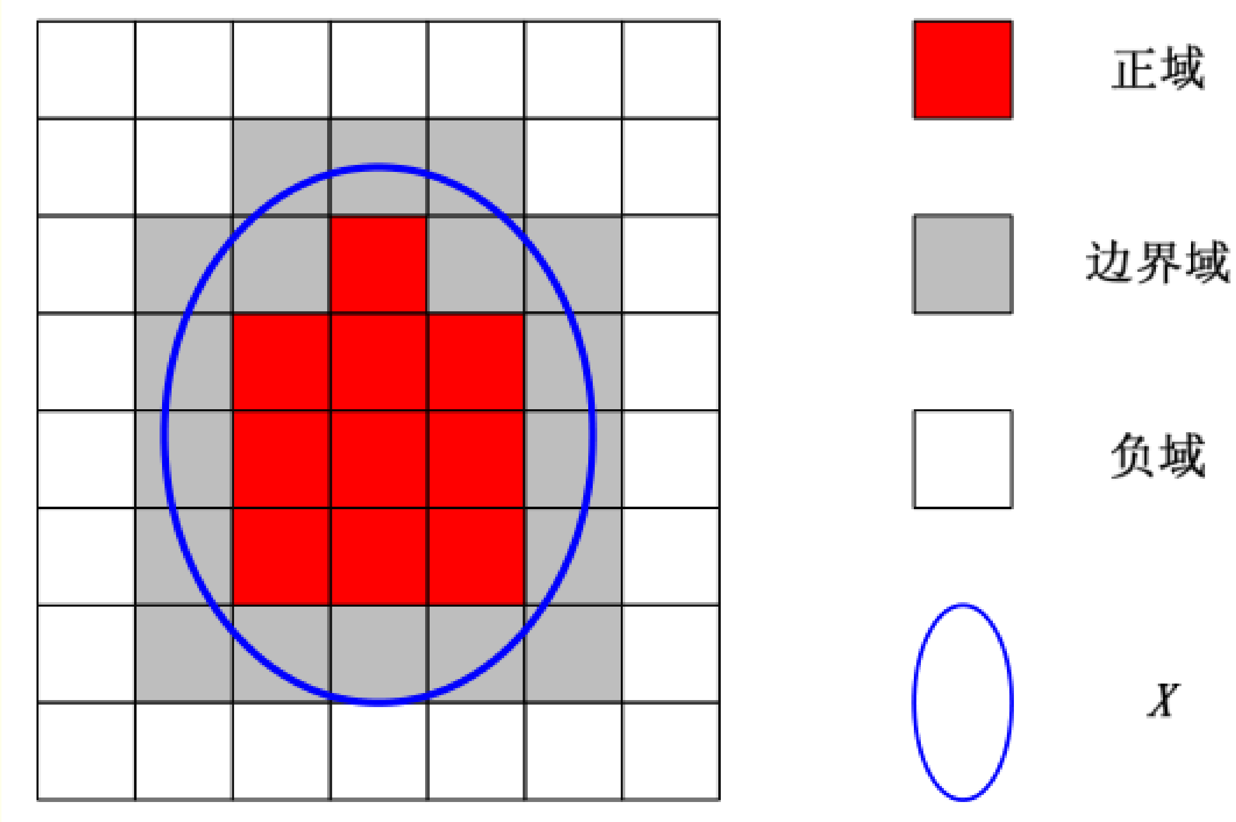
$$\underline{R}(X)=\{x \mid [x]_R \subseteq X\}$$

$$\overline{R}(X)=\{x \mid [x]_R \cap X \neq \phi\}$$

- 集合X的R边界域：$BN_R(X) = \overline{R}X - \underline{R}X$
- 集合X的R正域：$POS_R(X) = \underline{R}X$
- 集合X的R负域：$NEG_R(X) = U - \overline{R}X$

集合X是R精确集，当且仅当上下近似相等。

集合X是R粗糙集，当且仅当上下近似不相等。



粗糙集中相关概念的示意图

【例】对于如表9.4所示的决策表T_1，C={头疼，肌肉疼，体温}，D={流感}，求$POS_C(D)$。

U	条件属性			决策属性
	头疼	肌肉疼	体温	流感
1	是	是	正常	否
2	是	是	高	是
3	是	是	很高	是
4	否	是	正常	否
5	否	否	高	否
6	否	是	很高	是

$U/D=\{\{1, 4, 5\}, \{2, 3, 6\}\}$

设$X_1=\{1, 4, 5\}$，$X_2=\{2, 3, 6\}$

$U/C=\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$

$$\underline{R_C}(X_1)=\{1,4,5\} \qquad \underline{R_C}(X_2)=\{2,3,6\}$$

$$POS_C(D)=\underline{R_C}(X_1)\cup\underline{R_C}(X_2)=\{1,2,3,4,5,6\}$$

U	条件属性			决策属性
	头疼	肌肉疼	体温	流感
1	是	是	正常	否
2	是	是	高	是
3	是	是	很高	是
4	否	是	正常	否
5	否	否	高	否
6	否	是	很高	是

Pawlak粗糙集的稳健性分析

从粗糙集的定义可以看出，当一个等价类中所有的样本都属于*Y*时，这个等价类才能被划分到*Y*的正域。如果这个等价类中只有一个样本不属于集合*Y*，那么整个等价类都将被划分到*Y*的边界域。这样就直接影响了*Y*的近似精度。因此，可以得出Pawlak粗糙集的下上近似对噪声是十分敏感的。

2. 邻域粗糙集

定义 给定一个邻域近似空间< U,NR>和X ⊆ U，X在邻域近似空间<U,NR>上的下近似和上近似分别定义为

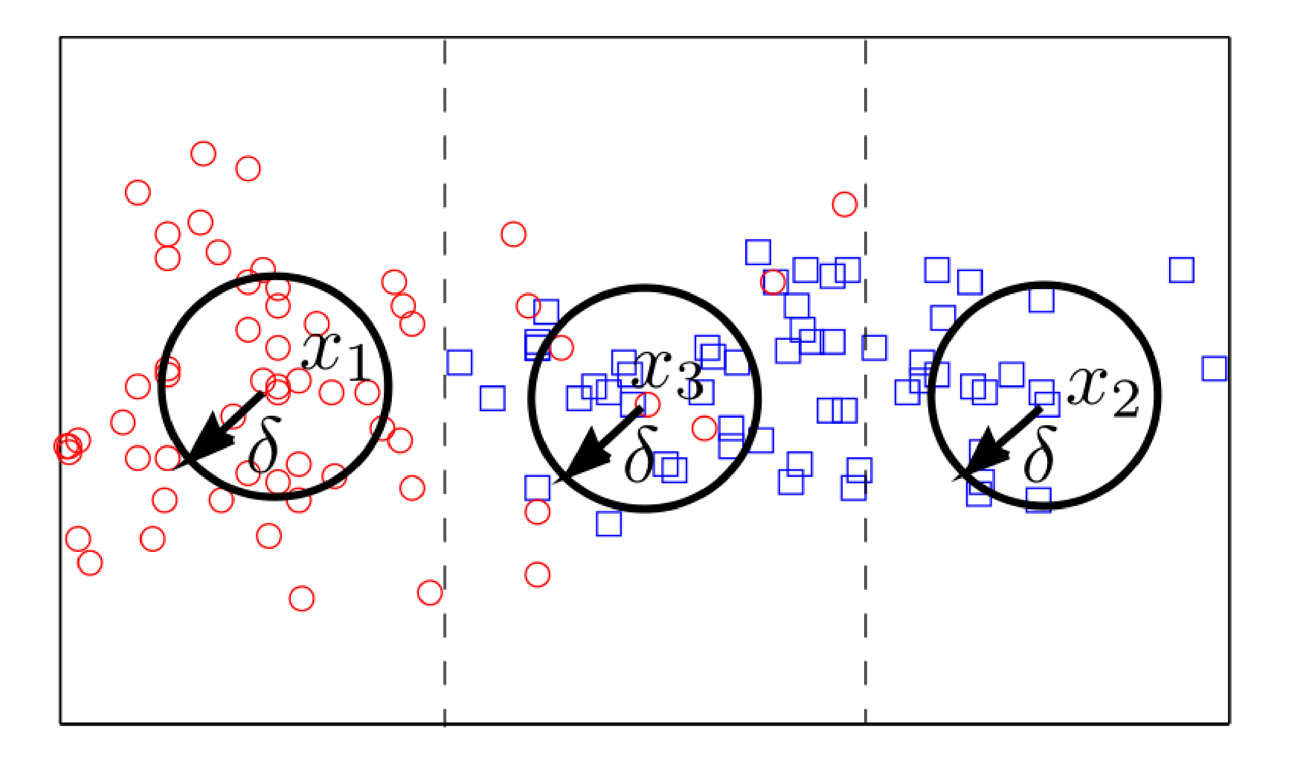
$$\underline{NR}X = \{x_i | \delta(x_i) \subseteq X, x_i \in U\},$$
$$\overline{NR}X = \{x_i | \delta(x_i) \cap X \neq \emptyset, x_i \in U\}.$$

定义：给定一个邻域决策系统<U,NR,D>，D将U划分为P个等价类：$X_1, X_2, \ldots, X_p$。$B \subseteq A$生成U上的邻域关系NR_B，则决策D关于B的邻域下近似和上近似分别为

$$\begin{cases} \underline{NR_B}D = \{\underline{NR_B}X_1, \underline{NR_B}X_2, \ldots, \underline{NR_B}X_P\}, \\ \overline{NR_B}D = \{\overline{NR_B}X_1, \overline{NR_B}X_2, \ldots, \overline{NR_B}X_P\}. \end{cases}$$

其中

$$\begin{cases} \underline{NR_B}X = \{x_i \mid \delta_B(x_i) \subseteq X, x_i \in U\}, \\ \overline{NR_B}X = \{x_i \mid \delta_B(x_i) \cap X \neq \emptyset, x_i \in U\}, \end{cases}$$



领域粗糙集示意图

3. 模糊粗糙集

在Pawlak粗糙集理论中，等价关系是最基本的概念。对于模糊粗糙集而言，模糊等价关系是最基本的概念。设U是非空论域，论域U上的模糊关系R称作模糊等价关系，如果R满足

(1) 自反性：$R(x,x)=1$,

(2) 对称性：$R(x,y)=R(y,x)$,

(3) 传递性：$R(x,y)\geq \sup\min_{z\in U}\{R(x,z),R(z,y)\}$,

对于任意的$x\in U$，x的模糊等价类定义为$[x]_R(y)=R(x,y),y\in U$。

定义 设U是非空论域，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 ∈ F(U)的下、上近似定义为

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

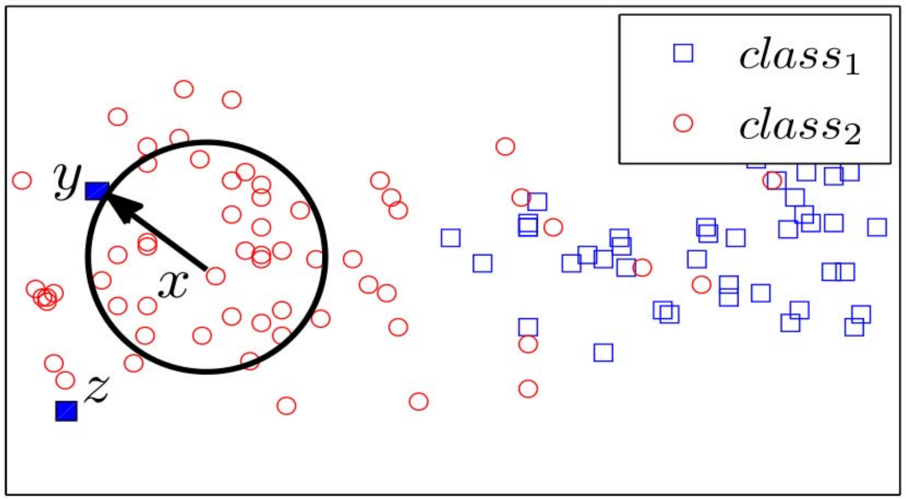
定义 设U是非空论域，R是U上的模糊等价关系，模糊集A $\in$ F(U)的上下近似定义为

$$\overline{R_T}A(x) = \sup_{u \in U} T(R(x,u), A(u));$$
$$\underline{R_S}A(x) = \inf_{u \in U} S(N(R(x,u), A(u));$$
$$\overline{R_\sigma}A(x) = \sup_{u \in U} \sigma(N(R(x,u)), A(u));$$
$$\underline{R_\vartheta}A(x) = \inf_{u \in U} \vartheta(R(x,u), A(u)).$$

模糊粗糙集的稳健性分析

令决策属性把样本分为k个子集分别为D_1, $D_2, \dots, D_k$。如果令T(x,y)=min(x,y),S(x,y)=max(x,y),则

$$\begin{cases} \underline{R}D_i(x) = \inf\limits_{y \notin D_i} \{1 - R(x,y)\}, \\ \overline{R}D_i(x) = \sup\limits_{y \in D_i} \{R(x,y)\}. \end{cases}$$



三、稳健粗糙集模型稳健原理分析

- 变精度粗糙集
- 邻域一致度
- 模糊变精度粗糙集
- 软模糊粗糙集
- 基于SMEB的模糊粗糙集
- 基于稳健统计量的模糊粗糙集

1.变精度粗糙集

设(U,R)为近似空间，论域U为非空有限集合，R为U上的等价关系，$U/R=\{E_1,E_2,...,E_n\}$为R的等价类构成的集合。对于$X\subseteq U$，X的β下近似定义为

$$\underline{R}_\beta X=\cup\{E\in U/R|e(E,X)\le\beta\},$$

上近似定义为

$$\overline{R}_\beta X=\cup\{E\in U/R|e(E,X)<1-\beta\}.$$

随后，Katzberg和Ziarko提出了另一种变精度粗糙集模型，该模型引入$0\le l<u\le 1$来代替β。对于经典集合$X\subseteq U$，X的u-下近似和l-上近似定义为

$$\begin{cases}\underline{R}_u X=\{x\in U|e([x]_R,X)\le 1-u\},\\ \overline{R}_l X=\{x\in U|e([x]_R,X)<1-l\}.\end{cases}$$

2. 邻域一致度

定义 2.6 给定邻域决策系统$NDT=<U,C,D>$，$x_i\in U$，$\delta(x_i)$是样本x_i在邻域近似空间的邻域。$P(\omega_j|\delta(x_i))$，$j=1,2,...,c$，是该邻域内ω_j类的概率，那么定义x_i的邻域决策函数为

$$ND(x_i)=w_j, P(\omega_j|\delta(x_i))=\max_j P(\omega_j|\delta(x_i)),$$

其中$P(\omega_j|\delta(x_i)=n_j/n$，$n$是邻域内样本的数量，$n_j$是第$j$类样本的数量。

误分类损失函数为

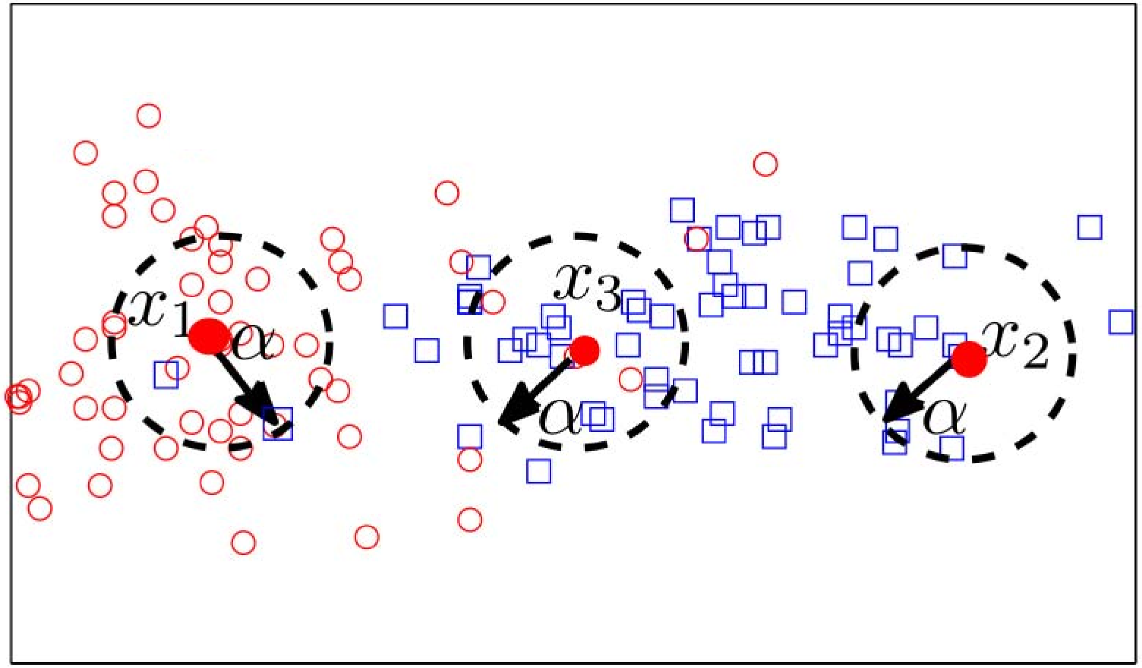
$$\lambda(\omega(x_i)|ND(x_i))=\begin{cases}0, \omega(x_i)=ND(x_i)\\ 1, \omega(x_i)\neq ND(x_i)\end{cases},$$

其中，$\omega(x_i)$是x_i的真实类别。

3. 模糊变精度粗糙集

定义 2.10 给定一个模糊决策系统$< U, C, D >$和一个模糊集合$A \in F(U)$，D是决策。对于给定的变精度参数$\alpha \in (0,1]$ 和任意$x \in U$，模糊变精度粗糙集的下、上近似定义如下 ：

$$\begin{aligned}
\underline{R_{\vartheta}}_{\alpha} A(x) &= \inf_{A(u)\le\alpha} \vartheta(R(x,u),\alpha) \wedge \inf_{A(u)>\alpha} \vartheta(R(x,u),A(u)),\\
\overline{R_T}_{\alpha} A(x) &= \sup_{A(u)\ge N(\alpha)} T(R(x,u),N(\alpha)) \vee \sup_{A(u)<N(\alpha)} T(R(x,u),N(\alpha)),\\
\underline{R_S}_{\alpha} A(x) &= \inf_{A(u)\le\alpha} S(N(R(x,u)),\alpha) \wedge \inf_{A(u)>\alpha} S(N(R(x,u)),A(u)),\\
\overline{R_{\delta}}_{\alpha} A(x) &= \sup_{A(u)\ge N(\alpha)} \delta(N(R(x,u)),N(\alpha)) \vee \sup_{A(u)<N(\alpha)} \delta(N(R(x,u)),N(\alpha)).
\end{aligned}$$



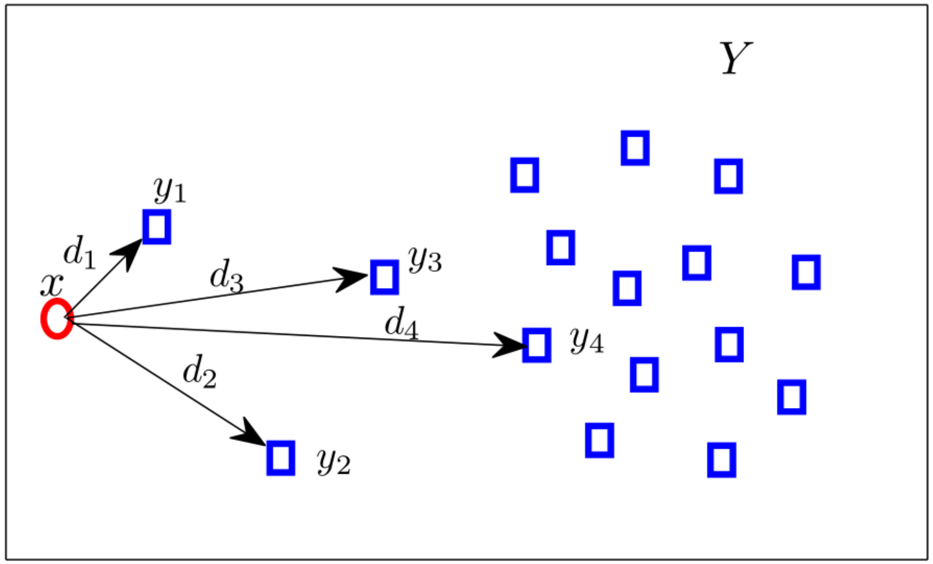
模糊变精度粗糙集的下近似

4. 软模糊粗糙集

定义　给定一个点集$Y=\{y_1,y_2,...,y_n\}$和集合外一点x，x到Y的软距离定义为

$$SD(x,Y)=\arg_{d(x,y_i)}\max\{d(x,y_i)-\beta\cdot m_i, y_i\in Y, i=1,2,...,n\},$$

其中，$d(\cdot,\cdot)$是距离函数，β是惩罚因子，m_i是集合Y中满足条件$d(x,y_j)<d(x,y_i)(j=1,2,...,n)$的点数。



点到集合的距离

定义 令U是一个非空论域，R是U上的模糊等价关系，$F(U)$是U的模糊幂集。对于$A \in F(U)$，软模糊粗糙集的下、上近似定义为

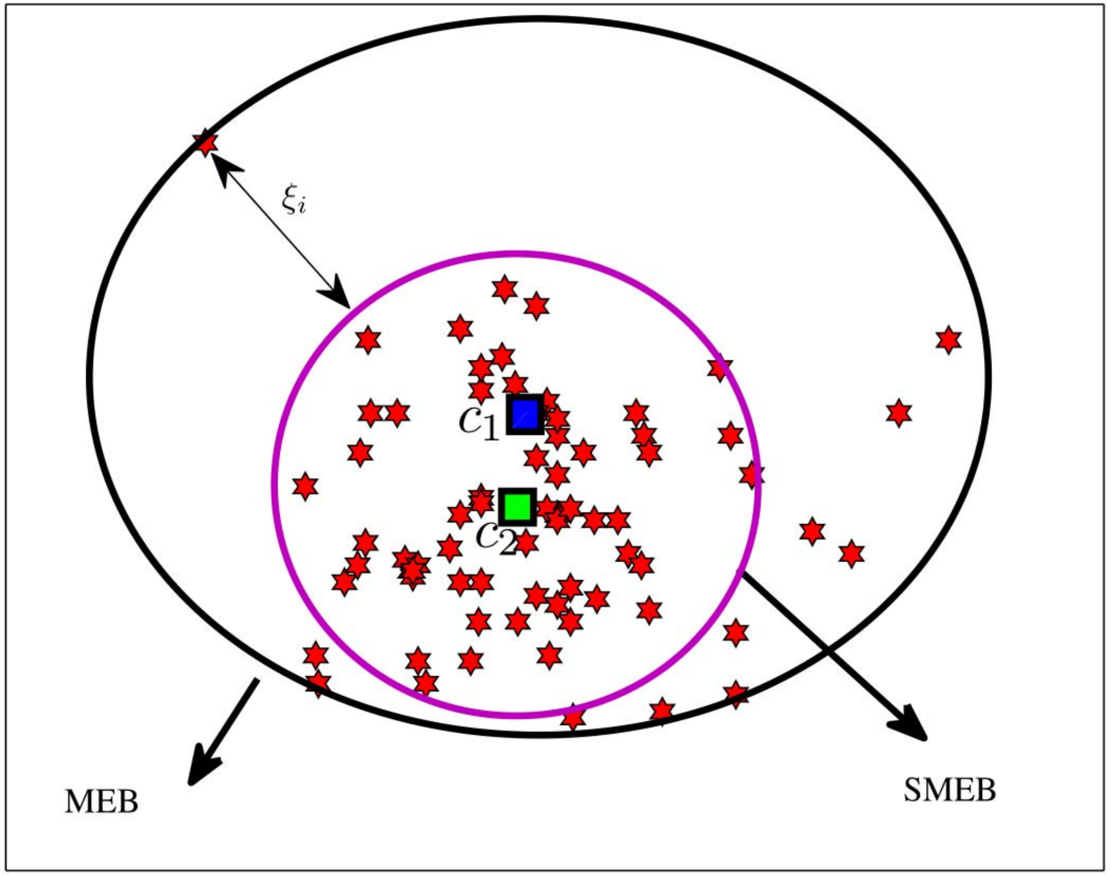
$$\begin{cases} \underline{R^S}(A)(x) = 1 - R(x, \arg_y \sup_{A(y) \le A(y_L)} \{1 - R(x, y) - \beta m_{Y_L}\}), \\ \overline{R^S}(A)(x) = R(x, \arg_y \inf_{A(y) \ge A(y_U)} \{R(x, y) + \beta n_{Y_U}\}). \end{cases}$$

其中，

$$\begin{cases} Y_L = \{y | A(y) \le A(y_L), y \in U\}, y_L = \arg_y \inf_{y \in U} \max\{1 - R(x, y), A(y)\}, \\ Y_U = \{y | A(y) \ge A(y_U), y \in U\}, y_U = \arg_y \sup_{y \in U} \min\{R(x, y), A(y)\}. \end{cases}$$

β是惩罚因子，m_{Y_L}是在计算下近似时被忽略的样本数量，n_{Y_U}是在计算上近似时被忽略的样本数量。

5. 基于SMEB的模糊粗糙集



最小超球与软最小超球

定义 令U是一个非空集合，R是U上的模糊等价关系，且$F(U)$是U的模糊幂集。对于模糊集$A \in F(U)$，基于软最小超球的模糊粗糙集的上下近似算子定义为

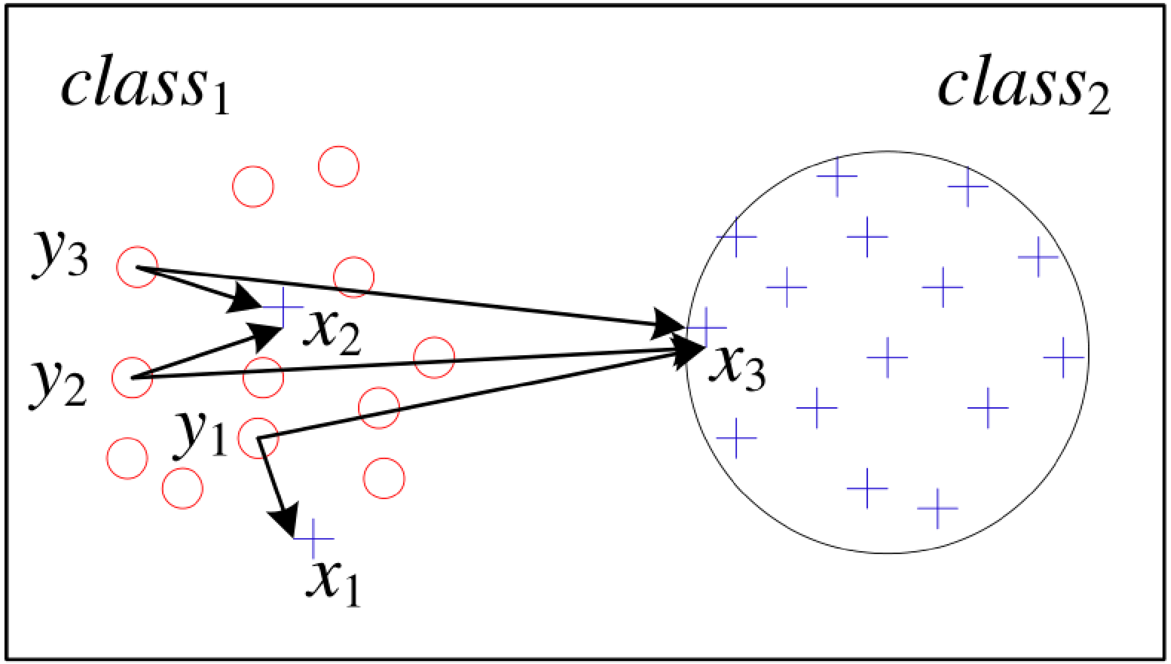
$$\begin{cases} \overline{R_T}^R A(x) = \sup_{y \in U} T(R(x,y), \mathrm{SMEB}_{Y_U}(c,r)(y)), \\ \underline{R_S}^R A(x) = \inf_{y \in U} S(N(R(x,y)), N(\mathrm{SMEB}_{Y_L}(c,r)(y))), \end{cases}$$

其中，

$$\mathrm{SMEB}_*(c,r)(y) = \begin{cases} 1, \| \phi(y) - c \| \le r, \\ 0, \| \phi(y) - c \| > r, \end{cases}$$

和

$$\begin{cases} Y_U = \{y \in U | A(y) \ge A(y_U)\}, y_U = \arg\sup_{y \in U} T(R(x,y), A(y)), \\ Y_L = \{y \in U | A(y) \le A(y_L)\}, y_L = \arg\inf_{y \in U} S(N(R(x,y)), A(y)). \end{cases}$$



基于软最小超球的模糊粗糙集的下近似图解

5. 基于稳健统计量的模糊粗糙集

定义 对于随机变量$X=\{x_1,x_2,...,x_n\}$，其中X中的元素是按照升序排列的，X的k-trimmed最小值和最大值分别为x_{k+1}和x_{n-k}，X的k-mean最小值和最大值分别为$\sum_{i=1}^{k} x_i/k$和$\sum_{i=n-k+1}^{n} x_i/k$，X的k-median最小值和最大值分别为$\{x_1,...,x_k\}$的中位数和$\{x_{n-k+1},...,x_n\}$的中位数。

定义 对于决策表$DT=<U,C,D>$，R是由特征子集$B\subseteq C$诱导出的模糊相似关系，且$R(x,y)$随着距离$\Delta(x,y)$的增加单调减小。令d_i是第i类样本的集合，对于$\forall x\in d_i$，下、上近似算子分别定义为

$$\underline{R_S}_{k-trimmed}d_i(x)=\min_{y\notin d_i\ k-trimmed}1-R(x,y);$$
$$\overline{R_T}_{k-trimmed}d_i(x)=\max_{y\in d_i\ k-trimmed}R(x,y);$$
$$\underline{R_\vartheta}_{k-trimmed}d_i(x)=\min_{y\notin d_i\ k-trimmed}\sqrt{1-R^2(x,y)};$$
$$\overline{R_\sigma}_{k-trimmed}d_i(x)=\max_{y\in d_i\ k-trimmed}1-\sqrt{1-R^2(x,y)}.$$

$$\underline{R_S}_{k-mean}d_i(x)=\min_{y\notin d_i\ k-mean}1-R(x,y);$$
$$\overline{R_T}_{k-mean}d_i(x)=\max_{y\in d_i\ k-mean}R(x,y);$$
$$\underline{R_\vartheta}_{k-mean}d_i(x)=\min_{y\notin d_i\ k-mean}\sqrt{1-R^2(x,y)};$$
$$\overline{R_\sigma}_{k-mean}d_i(x)=\max_{y\in d_i\ k-mean}1-\sqrt{1-R^2(x,y)}.$$

$$\underline{R_S}_{k-median}d_i(x)=\min_{y\notin d_i\ k-median}1-R(x,y);$$
$$\overline{R_T}_{k-median}d_i(x)=\max_{y\in d_i\ k-median}R(x,y);$$
$$\underline{R_\vartheta}_{k-median}d_i(x)=\min_{y\notin d_i\ k-median}\sqrt{1-R^2(x,y)};$$
$$\overline{R_\sigma}_{k-median}d_i(x)=\max_{y\in d_i\ k-median}1-\sqrt{1-R^2(x,y)}.$$

四、粗糙集理论的应用

- 鲁棒特征选择
- 鲁棒分类
- 鲁棒预测

基于Pawlak属性重要度的属性约简算法

为了度量某个条件属性相对于决策属性的重要性，或者说是该属性在整个决策表中的作用，粗糙集理论采用的方式是依次从决策表中删除每个属性，然后观察该属性被删除之后，整个决策表的分类能力有没有发生变化，发生变化的幅度是多大。

如果变化越大，则说明，所删除的属性对原始决策表就越重要，反之，如果变化很小，说明所删除属性对原始决策表而言就不是很重要。

定义 设$T=(U, C, D)$是一个协调的决策表，$\forall a\in C$，该属性对条件属性全集C相对于决策属性D的重要度定义为：

$$sig(a,C;D)=r_C(D)-r_{C-\{a\}}(D)=\frac{|POS_C(D)|-|POS_{C-\{a\}}(D)|}{|U|}$$

$\forall B\subseteq C$，$\forall a\in C-B$，条件属性a对条件属性集B相对于决策属性D的重要度定义为：

$$sig(a,B;D)=r_{B\cup\{a\}}(D)-r_B(D)=\frac{|POS_{B\cup\{a\}}(D)|-|POS_B(D)|}{|U|}$$

属性约简算法步骤：

1、计算C相对于D的核 B=$CORE_C(D)$;

2、若$POS_B(D)=POS_C(D)$则算法结束；若$POS_B(D)\neq POS_C(D)$，则分别计算每个非核属性对核集B相对于D的重要度，选择最大重要度的属性加入核集B。

3、重复步骤2。

【例】对于如表9.9所示的决策表T_4，采用基于Pawlak属性重要度的属性约简算法求相对属性约简集B。求解过程如下：

① 计算核。

$U/C=\{\{1\},\{2\},\{3\},\{4\}\}$

$U/D=\{\{1,2\},\{3,4\}\}$

$U/(C-\{a\})=U/\{b,c,d\}=\{\{1\},\{2\},\{3\},\{4\}\}$

$U/(C-\{b\})=U/\{a,c,d\}=\{\{1\},\{2\},\{3,4\}\}$

$U/(C-\{c\})=U/\{a,b,d\}=\{\{1,4\},\{2\},\{3\}\}$

$U/(C-\{d\})=U/\{a,b,c\}=\{\{1\},\{2\},\{3,4\}\}$

$POS_C(D)=\{1,2,3,4\}$

$POS_{C-\{a\}}(D)=\{1,2,3,4\}=POS_C(D$

$POS_{C-\{b\}}(D)=\{1,2,3,4\}=POS_C(D)$

$POS_{C-\{c\}}(D)=\{2,3\}\neq POS_C(D)$

$POS_{C-\{d\}}(D)=\{1,2,3,4\}=POS_C(D))$

所以条件属性c是决策表的核值属性，即$CORE_C(D)=\{c\}$。

U	条件属性				决策属性
	a	***b***	***c***	***d***	***e***
1	0	0	0	0	0
2	1	0	1	1	0
3	0	1	1	0	1
4	0	0	1	0	1

② 令$B=CORE_C(D)=\{c\}$，$POS_B(D)=\{1\}\neq POS_C(D)$，对于$C-B=\{a, b, d\}$中的每个属性$c_i$有其重要度：

$U/D=\{\{1, 2\}, \{3, 4\}\}$ $U/B=\{\{1\}, \{2, 3, 4\}\}$

$U/\{a, c\}=\{\{1\}, \{2\}, \{3, 4\}\}$ $U/\{b, c\}=\{\{1\}, \{3\}, \{2, 4\}\}$

$U/\{d, c\}=\{\{1\}, \{2\}, \{3, 4\}\}$ $POS_B(D)=\{1\}$

$POS_{B\cup\{a\}}(D)=POS_{\{a, c\}}(D)=\{1, 2, 3, 4\}$

$POS_{B\cup\{b\}}(D)=POS_{\{b, c\}}(D)=\{1, 3\}$

$POS_{B\cup\{d\}}(D)=POS_{\{d, c\}}(D)=\{1, 2, 3, 4\}$

$$sig(a,B;D)=\frac{|POS_{B\cup\{a\}}(D)|-|POS_B(D)|}{|U|}=\frac{4-1}{4}=\frac{3}{4}$$

$$sig(b,B;D)=\frac{|POS_{B\cup\{b\}}(D)|-|POS_B(D)|}{|U|}=\frac{2-1}{4}=\frac{1}{4}$$

$$\mathrm{sig}(d, B; D)=\frac{|\mathrm{POS}_{B\cup\{d\}}(D)|-|\mathrm{POS}_B(D)|}{|U|}=\frac{4-1}{4}=\frac{3}{4}$$

取c_m=arg max sig(c_i，B；D)，可选的属性有a和d

B={a，c}或{d，c}，

由于$POS_{\{a, c\}}(D)=POS_C(D)$，$POS_{\{d, c\}}(D)=POS_C(D)$

算法结束。

输出的约简为B={a，c}或{d，c}。

总　结

- 大数据时代，低质大数据严重影响数据分析的结果，鲁棒建模十分必要。
- 理论研究服务于工程实际应用，建立鲁棒模型避免造成严重的后果。
- 鲁棒粗糙集建模研究成果广泛应用于各个领域。