S-粗集与数据挖掘单位圆特征

史开泉
(山东大学数学与系统科学学院 济南 250100)

摘 要 给出单向S-粗集(one direction singular rough sets)、单向S-粗集对偶(dual of one direction singular rough sets)的结构。单向S-粗集与单向S-粗集对偶是改进Z. Pawlak粗集得到的,单向S-粗集与单向S-粗集对偶具有动态特性。给出单向S-粗集、单向S-粗集对偶与Z. Pawlak粗集的关系。S-粗集具有三类形式:单向S-粗集、单向S-粗集对偶、双向S-粗集,利用单向S-粗集、单向S-粗集对偶,给出数据内挖掘、数据外挖掘概念,给出数据内挖掘的外同心圆定理、数据外挖掘的内同心圆定理,并给出其应用。S-粗集是粗集理论与应用研究的新分支。

关键词 单向S-粗集,单向S-粗集对偶,数据内挖掘,数据外挖掘,单位圆,外同心圆定理,内同心圆定理,应用

S-rough Sets and Characteristics of Data Mining Unit Circle

SHI Kai-quan
(School of Mathematics and System Sciences, Shandong University, Jinan 250100, China)

Abstract One direction singular rough sets and the structure of dual of one direction singular rough sets were given. One direction singular rough sets and dual of one direction singular rough sets come from improved Z. Pawlak rough sets. One direction singular rough sets and dual of one direction singular rough sets have dynamic characteristics. The relation between one direction singular rough sets, dual of one direction singular rough sets and Z. Pawlak rough sets was also proposed. S-rough sets have three types of forms: one direction singular rough sets, dual of one direction singular rough sets and two directions singular rough sets. Based on one direction singular rough sets, dual of one direction singular rough sets, the concepts of data internal-mining, data outer-mining and their Outer concentric circle theorem, Internal concentric circle theorem were proposed, and the applications were given. S-rough sets is a new branch of Rough sets theory and applied research.

Keywords One direction singular rough sets, Dual of one direction singular rough sets, Data internal-mining, Data outer-mining, Unit circle, Outer concentric circle theorem, Internal concentric circle theorem, Application

1 引言

一个例子:五一长假前夕,我的学生们(博士,硕士)x_1,x_2,x_3,x_4,x_5,给我提出一个问题:"五一期间,我们想去上海的几所高校做一次学术访问,也借此机会在上海放松几天,老师可以否?"我回答学生们:"可以,由济南—上海的往返车费与住宿费,可以从我的研究经费中支付。"从这个简单的对话中,想到一个数学概念:因为 x_1,x_2,x_3,x_4,x_5,都在 α=上海下车,x_1-x_5 关于 α 构成等价类$[x]=\{x_1,x_2,x_3,x_4,x_5\}$($x_1$,$x_2$,$x_3$,$x_4$,$x_5$,关于 α 满足自反性,对称性,传递性)。

1° 学生 x_6,x_7 原本没有去上海做学术访问的计划,当听到老师支付车费与住宿费,还可以在上海放松几天时:x_6,x_7 改变了计划,x_6,x_7 随同 x_1-x_5 乘上济南赴上海的火车(x_6,x_7 在 x_1-x_5 的带领下,在火车上临时补车票)。因此,$[x]=\{x_1,x_2,x_3,x_4,x_5\}$变成$[x]^*=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7\}$;或者,$[x]^*$是在$[x]$内进行元素补充得到的。在由济南赴上海的火车上(济南去上海的火车途经南京),原籍南京的学生 x_3 接到家中电话,要 x_3 立即回家,x_3 在南京站下了火车。因此$[x]=\{x_1,x_2,x_3,x_4,x_5\}$变成$[x]^\circ=\{x_1,x_2,x_4,x_5\}$;或者,$[x]^\circ$是在$[x]$内进行元素删除得到的。容易看到:等价类$[x]$具有了动态特性:$[x]$变成$[x]^*$,$[x]$变成$[x]^\circ$。

2° 在计算机科学领域中,数据库是一个被人们熟悉的概念,用等价类$[x]$表示。人们常常要对数据库进行刷新:$[x]$内的一些过时的数据要从$[x]$内删除,一些新的数据要补充到$[x]$内;数据库的这些司空见惯的特性,没有引起人们的兴趣,也没有人对这些特性给出开发及应用研究。

1°,2° 给出一个事实:等价类$[x]$具有了动态特性;$[x]$的动态特性催生了作者改进Z. Pawlak粗集,提出新粗集的想法。1982年Z. Pawlak教授提出粗集(Rough sets)[1],给出粗集的数学结构:$(R_-(X),R^-(X))$;这里:$R_-(X)=\bigcup[x]=\{x|x\in U,[x]\subseteq X\}$,$R^-(X)=\bigcup[x]=\{x|x\in U,[x]\cap X\neq\phi\}$,分别是 $X\subset U$ 的下近似、上近似;$[x]$是R-等价类。Z. Pawlak的这一杰出的原创性研究获得了应用,得到了人们的关注与认可。"近似逼近"是Z. Pawlak粗集存在的"基石与根本";或者用 $R_-(X)$与 $R^-(X)$共同"逼近"集合 $X\subset U$,$X\subset U$是U上的有限集。"等价类$[x]$"是Z. Pawlak粗集定义中的

到稿日期:2010-03-09 本文受山东省自然科学基金项目(Y2007G08)资助。

史开泉 教授,博士生导师,主要研究领域为粗集理论与应用、信息系统与信息识别理论与应用,E-mail:shikq@sdu.edu.cn。

"核心概念"。应当看到：Z. Pawlak 粗集中的等价类$[x]$不具有"动态特性"，$[x]$具有"静态特性"；显然，"静态特性"限定了 Z. Pawlak 粗集的应用空间，限制了 Z. Pawlak 粗集的广泛应用，特别是在动态信息系统中。

2002 年，文献[2,3]改进了 Z. Pawlak 粗集，提出 S-粗集(Singular rough sets)；S-粗集具有三类形式：单向 S-粗集(one direction singular rough sets)、单向 S-粗集对偶(dual of one direction singular rough sets)与双向 S-粗集(two directions singular rough sets)。本文中只给出单向 S-粗集、单向 S-粗集对偶的结构与应用；双向 S-粗集的结构见文献[4-6,21,22]。

2 单向 S-粗集与单向 S-粗集对偶

文献[2-6,21,22]给出约定：U 是有限元素论域，$[x]$是 U 上的元素等价类，V 是有限属性论域，$F=\{f_1,f_2,\cdots,f_m\}$，$\overline{F}=\{\overline{f}_1,\overline{f}_2,\cdots,\overline{f}_n\}$是元素迁移族；$f\in F$，$\overline{f}\in\overline{F}$是元素迁移，$f\in F$ 的特征是：$u\in U,u\overline{\in}X,f\in F$ 把 u 变成 $f(u)=x'\in X$；$\overline{f}\in\overline{F}$ 的特征是：$\exists x\in X,\overline{f}\in\overline{F}$ 把 x 变成 $\overline{f}(x)=u\overline{\in}X$。

单向 S-粗集

称 $X^o\subset U$ 是 X 的一个单向 S-集合(one direction singular sets)，如果

$$X^o=X\cup\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\} \tag{1}$$

称 X^f 是 $X\subset U$ 的 f-扩张，而且

$$X^f=\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\} \tag{2}$$

式(1)，式(2)中的集合 X 是 Z. Pawlak 粗集$(R_-(X),R^-(X))$中的集合[1]，$X\subset U$。

这里指出："扩张"是指在 $X\subset U$ 内补充元素，使得 card$(X)\leqslant$card(X^o)，card=cardinal number。

称$(R,F)_o(X^o)$是 $X^o\subset U$ 的下近似，而且

$$(R,F)_o(X^o)=\cup[x]=\{x|x\in U,[x]\subseteq X^o\} \tag{3}$$

称$(R,F)^o(X^o)$是 $X^o\subset U$ 的上近似，而且

$$(R,F)^o(X^o)=\cup[x]=\{x|x\in U,[x]\cap X^o\neq\phi\} \tag{4}$$

式中，$F\neq\phi$。

由$(R,F)_o(X^o)$，$(R,F)^o(X^o)$构成的集合对，而且

$$((R,F)_o(X^o),(R,F)^o(X^o)) \tag{5}$$

称作 $X^o\subset U$ 的单向 S-粗集(one direction singular rough sets)。

称 Bnr(X^o)是 $X^o\subset U$ 的 R-边界，而且

$$\mathrm{Bnr}(X^o)=(R,F)^o(X^o)-(R,F)_o(X^o) \tag{6}$$

称 As(X^o)是单向 S-粗集生成的副集合(assistant set)，而且

$$\mathrm{As}(X^o)=\{x|u\in U,u\overline{\in}X,f(u)=x\widetilde{\in}X\} \tag{7}$$

这里特别给出说明：

1°. 式(2)中 $X^f=\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}$是被补充到 X 内的新元素构成的集合，$\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}$与被补充新元素之前的 X 满足：$\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}\cap X=\phi$。专家在审查我的论文时曾给出这样的结论："因为$\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}\subseteq X$，所以式(1)成为：$X^o=X\cup\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}=X$，因此式(1)无意义。"显然，这是一个在没有看懂式(1)具体意义的情况下，给出的"瞎掰"评论。通常的 $A\subseteq B$，是指 $A=\{x_1,x_2,x_3\}\subseteq\{x_1,x_2,x_3,x_4\}=B$，从何而来的$\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}\subseteq X$？$X^o=X\cup\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}=X$ 可以得到这样的类比例子："甲第一次还借款 x 给乙；甲第二次还借款 y 给乙；乙却说借款 y 没有还"；显然，乙是人们常说的一句话："乙是一个无赖"。事实上，式(1)的结构与计算机内存储器 $T=T+1$的结构相似，$T=T+1$ 具有动态特性，式(1)也具有动态特性。专家给出的令人啼笑皆非的评论："$X^o=X\cup\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}=X$"是因为用"静态"的思维方式去认识式(1)的结构，不知道式(1)所表示的含义，或者不具备计算机内存储器结构的基本常识。式(1)的动态特征是：取 $u_1\in U,u_1\overline{\in}X,f\in F$ 把 u_1 变成 $f(u_1)=x_1'\in X$，式(1)变成：$X_1^o=X\cup\{u_1|u_1\in U,u_1\overline{\in}X,f(u_1)=x_1'\in X\}=X\cup\{x_1'\}=\{X,x_1'\}$。令 $X=X_1^o$，取 $u_2,u_3\in U,u_2,u_3\overline{\in}X,f\in F$ 把 u_2,u_3 变成 $f(u_2)=x_2'\in X,f(u_3)=x_3'\in X$，式(1)变成 $X_2^o=X_1^o\cup\{x_2',x_3'\}=\{X_1^o,x_2',x_3'\}=\{X,x_1',x_2',x_3'\}$。令 $X=X_2^o$，取 $u_4\in U,u_4\overline{\in}X,f\in F$ 把 u_4 变成 $f(u_4)=x_4'\in X$，式(1)变成 $X_3^o=X_2^o\cup\{x_4'\}=\{X_1^o,x_2',x_3'\}\cup\{x_4'\}=\{X,x_1',x_2',x_3',x_4'\}$；如此等等。这个过程正是 $T=T+1$ 表现的存储信息的动态过程，从这个过程中，无论怎样也得不到"因为$\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}\subseteq X$，所以 $X^o=X\cup\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}=X$"这样的结论。式(1)中的并运算"$\cup$"已不是经典数学或离散数学中的并运算。

2°. 式(1)，式(2)表示不在 X 内的元素 u，通过元素迁移 $f\in F$ 把 u 变成 $f(u)=x'$，$f(u)=x'$进入 X 内，X 变成 X^o。显然，$f(u)$进入 X 是指进入某一个$[x]$内；因此，式(3)，式(4)等价类$[x]$都具有向外扩张的特征，或者式(3)，式(4)中的$[x]$具有了动态特征。应该特别指出：式(3)，式(4)中使用的符号$[x]$与 Z. Pawlak 粗集$(R_-(X)=\cup[x]=\{x|x\in U,[x]\subseteq X\},R^-(X)=\cup[x]=\{x|x\in U,[x]\cap X\neq\phi\})$使用的符号$[x]$相同，但意义却不一样：式(3)，式(4)中的$[x]$具有动态特性，Z. Pawlak 粗集中的$[x]$具有静态特性。式(3)，式(4)中的$[x]$与 Z. Pawlark 粗集中的$[x]$不是一个概念，尽管都使用了符号$[x]$。

3°. 式(7)的意义是：$u\in U$ 被 $f\in F$ 变成 $f(u)=x$，$f(u)$被不完全迁入到 X 内，因此在式(7)中使用了一个特别的记号"$\widetilde{\in}$"。被完全迁入到 X 内的元素 $f(u)=x'$的特征函数是 $\chi_X^{(f(u))}=1$；被不完全迁入到 X 内的元素 $f(u)=x'$的特征函数 $0<\chi_X^{(f(u))}<1$。

4°. $F=\{f_1,f_2,\cdots,f_m\}$是元素迁移族，$f\in F$ 是元素迁移，元素迁移 $f\in F$ 是一个变换，或者是一个函数。不同的应用问题，$f\in F$ 的形式是不同的。"元素迁移"的概念，通俗地讲：山东籍的学生到重庆去读大学，这个学生必须带上"户口迁移证"，才能成为重庆市的人(按中国的户籍管理制度)；显然"户口迁移证"具有元素迁移的意义。

1°－4° 的说明对于接收单向 S-粗集的结构式(1)－式(7)是重要的。

利用式(1)－式(7)及 Z. Pawlak 粗集[1]，容易得到：

定理 1 若元素迁移族 $F=\phi$，则单向 S-粗集与 Z. Pawlak 粗集满足

$$((R,F)_o(X^o),(R,F)^o(X^o))_{F=\phi}=(R_-(X),R^-(X)) \tag{8}$$

事实上，若 $F=\phi$，则式(2) $X^f=\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}=\phi$，式(1)成为 $X^o=X\cup\{u|u\in U,u\overline{\in}X,f(u)=x'\in$

$X\}=X$；式(3)成为$(R,F)_o(X^o)=\bigcup[x]=\{x|x\in U,[x]\subseteq X^o\}=\bigcup[x]=\{x|x\in U,[x]\subseteq X\}=R_-(X)$；式(4)成为$(R,F)^o(X^o)=\bigcup[x]=\{x|x\in U,[x]\cap X^o\neq\phi\}=\bigcup[x]=\{x|x\in U,[x]\cap X\neq\phi\}=R^-(X)$；式(6)成为$\mathrm{Bnr}(X^o)=(R,F)^o(X^o)-(R,F)_o(X^o)=R^-(X)-R_-(X)=\mathrm{Bnr}(X)$；式(7)成为$\mathrm{As}(X^o)=\phi$。因此，当$F=\phi$时，单向S-粗集回到了Z. Pawlak粗集的“原点”，单向S-粗集退化成Z. Pawlak粗集。

容易得到以下命题：

命题1 在静态-动态条件下，单向S-粗集是Z. Pawlak粗集的一般形式，Z. Pawlak粗集是单向S-粗集的特例。

命题2 副集$\mathrm{As}(X^o)=\phi$的单向S-粗集是Z. Pawlak粗集。

这里指出：1° 改进Z. Pawlak粗集，提出单向S-粗集，在单向S-粗集中没有改变Z. Pawlak粗集中的等价关系R。

2° 单向S-粗集中“singular”一词取自“奇异矩阵”(singular matrix)。一个矩阵A具有P个特性，因为对A进行某些变换，变换的结果使A丢掉了r个特性$r<p$，A变成奇异矩阵A^*。应当把r个特性返换给A，因此用“singular”表示返换的意义。我们遇到的有限集合X，大多具有动态特性；从一般意义上看，动态特性是集合X的“本性”；具有静态特性的集合X是少数。集合X的原本面貌是动态的，经典数学中给出的集合X，它具有静态特性，或者把动态特性丢掉了。因此，用“singular”表示把动态特性返换给X，得到单向S-集合$X^o=X\cup\{u|u\in U,u\overline{\in}X,f(u)=x'\in X\}$。一个通常的例子：重庆市火车站候车室内的旅客$x_1,x_2,\cdots,x_m$构成了旅客集合$X=\{x_1,x_2,\cdots,x_m\}$(因为$x_1,x_2,\cdots,x_m$都具有特征$\alpha=$火车票；$x_1,x_2,\cdots,x_m$构成集合$X$)，集合$X$是不是我们学过的课程“数学分析”，“高等代数”，“实变函数”，“离散数学”中的集合X？若是，则出现违背事实的问题：“火车已进站，具有去某地车票的人不准上火车；持有车票的新旅客不准进火车站候车室候车”。因为上述课程中的集合X具有静态特性，或者X内的元素$x_i\in X$不允许离开X，$x_i\overline{\in}X$；X外的元素y_i，$y_i\overline{\in}X$不允许进入X，$y_i\in X$。显然，重庆市火车站候车室内的旅客构成的集合$X=\{x_1,x_2,\cdots,x_m\}$已不是“数学分析”，“高等代数”，“实变函数”，“离散数学”中的集合X。旅客构成的集合$X=\{x_1,x_2,\cdots,x_m\}$具有了动态特性；因为：“火车站里有火车，车站里面有旅客，旅客手中提包裹，不是上车是下车”。

这里指出：为了简单，式(1)－式(7)给出单向S-粗集的结构，式(5)给出单向S-粗集的集合对表示形式。因为元素迁移$f\in F$的存在，单向S-粗集具有了动态特性；因此，单向S-粗集的一般性的表示形式是：

$$\{((R,F)_o(X^o)_i,(R,F)^o(X^o)_j)|i\in I,j\in J\}\tag{9}$$

式(9)是单向S-粗集的集合对族的形式，单向S-粗集是由一串集合对构成。式(9)中I,J是指标集合。

由定理1得到：

定理2 若元素迁移族$F=\phi$，则单向S-粗集与Z. Pawlak粗集满足

$$\{((R,F)_o(X^o)_i,(R,F)^o(X^o)_j)|i\in I,j\in J\}_{F=\phi}=(R_-(X),R^-(X))\tag{10}$$

式(10)指出：$F=\phi$的条件下，每一个单向S-粗集$((R,F)_o(X^o)_i,(R,F)^o(X^o)_j)$都回到了Z. Pawlak粗集$(R_-(X),R^-(X))$的“原点”。

单向S-粗集对偶

称$X'\subset U$是单向S-集合X^o的对偶，如果

$$X'=X-\{x|x\in X,\overline{f}(x)=u\overline{\in}X\}\tag{11}$$

称$X^{\overline{f}}$是$X\subset U$的$\overline{f}$-萎缩，而且

$$X^{\overline{f}}=\{x|x\in X,\overline{f}(x)=u\overline{\in}X\}\tag{12}$$

式(11)，式(12)中的集合X是Z. Pawlak粗集$(R_-(X),R^-(X))$中的集合X[1]。

这里指出：“萎缩”是指在$X\subset U$内删除部分元素，使得$\mathrm{card}(X')\leqslant\mathrm{card}(X)$。

称$(R,\overline{F})_o(X')$是$X'\subset U$的下近似，而且

$$(R,\overline{F})_o(X')=\bigcup[x]=\{x|x\in U,[x]\subseteq X'\}\tag{13}$$

称$(R,\overline{F})^o(X')$是$X'\subset U$的上近似，而且

$$(R,\overline{F})^o(X')=\bigcup[x]=\{x|x\in U,[x]\cap X'\neq\phi\}\tag{14}$$

式中，$X'\neq\phi,\overline{F}\neq\phi$。

由$(R,\overline{F})_o(X')$，$(R,\overline{F})^o(X')$构成的集合对，而且

$$((R,\overline{F})_o(X'),(R,\overline{F})^o(X'))\tag{15}$$

称作$X'\subset U$的单向S-粗集对偶(dual of one direction singular rough sets)。

称$\mathrm{Bnr}(X')$是$X'\subset U$的R-边界，而且

$$\mathrm{Bnr}(X')=(R,\overline{F})^o(X')-(R,\overline{F})_o(X')\tag{16}$$

称$\mathrm{As}(X')$是单向S-粗集生成的副集合(assistant set)，而且

$$\mathrm{As}(X')=\{x|x\in X,\overline{f}(x)=u\in X\}\tag{17}$$

式(17)的意义是：$x\in X$被$\overline{f}\in\tilde{F}$变成$\overline{f}(x)=u$，$\overline{f}(x)$不被完全迁出到X外，因此式(17)中使用了一个特别的记号“$\in$”。被完全迁出到X外的元素$\overline{f}(x)=u$的特征函数是$\chi_{\tilde{X}}^{(\overline{f}(x))}=-1$；被不完全迁出到$X$外的元素$\overline{f}(x)=u$的特征函数是$-1<\chi_{\tilde{X}}^{(\overline{f}(x))}<0$。

利用式(11)－式(17)及Z. Pawlak粗集[1]容易得到：

定理3 若元素迁移族$\overline{F}=\phi$，则单向S-粗集对偶与Z. Pawlak粗集满足

$$((R,\overline{F})_o(X'),(R,\overline{F})^o(X'))_{\overline{F}=\phi}=(R_-(X),R^-(X))\tag{18}$$

事实上，若$\overline{F}=\phi$，则式(12)$X^{\overline{f}}=\{x|x\in X,\overline{f}(x)=u\overline{\in}X\}=\phi$，式(11)变成$X'=X$；式(13)变成$(R,\overline{F})_o(X')=\bigcup[x]=\{x|x\in U,[x]\subseteq X'\}=\bigcup[x]=\{x|x\in U,[x]\subseteq X\}=R_-(X)$，式(14)变成$(R,\overline{F})^o(X')=\bigcup[x]=\{x|x\in U,[x]\cap X'\neq\phi\}=\bigcup[x]=\{x|x\in U,[x]\cap X\neq\phi\}=R^-(X)$；式(16)变成$\mathrm{Bnr}(X')=(R,\overline{F})^o(X')-(R,\overline{F})_o(X')=R^-(X)-R_-(X)=\mathrm{Bnr}(X)$；式(17)变成$As(X')=\phi$。因此，当$\overline{F}=\phi$时，单向S-粗集对偶回到了Z. Pawlak粗集的“原点”，单向S-粗集对偶退化成Z. Pawlak粗集。

容易得到以下命题：

命题3 在静态-动态条件下，单向S-粗集对偶是Z. Pawlak粗集的一般形式，Z. Pawlak粗集是单向S-粗集对偶的特例。

命题4 副集$\mathrm{As}(X')=\phi$的单向S-粗集对偶是Z. Pawlak粗集。

应当指出：改进Z. Pawlak粗集，提出单向S-粗集对偶，在单向S-粗集对偶中没有改变Z. Pawlak粗集中的等价关系R。

关于单向 S-粗集、单向 S-粗集对偶的更多特性、概念，见文献[21,22]。

图 1 给出了元素迁移的直观表示。

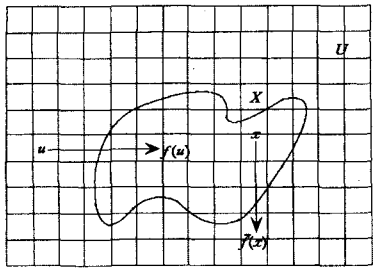

图 1　元素 $u\in U,u\in X$，元素迁移 $f\in F$ 把 u 变成 $f(u)=x'\in X$；元素 $x\in X$，元素迁移 $\bar{f}\in\bar{F}$ 把 x 变成 $\bar{f}(x)=u\in X$；X 是 U 上的有限元素集合；图中的每一个小方块表示等价类$[x]$。

单向 S-粗集、单向 S-粗集对偶的直观表示，见文献[21,22]。

这里指出：式(11)－式(17)给出单向 S-粗集对偶的结构，式(15)给出单向 S-粗集对偶的集合对表示形式。因为元素迁移 $\bar{f}\in\bar{F}$ 的存在，单向 S-粗集对偶具有了动态特性；所以，单向 S-粗集对偶的一般性的表示形式是：

$$\{((R,\bar{F})_o(X')_i,(R,\bar{F})^o(X')_j)\mid i\in I,j\in J\} \tag{19}$$

式(19)是单向 S-粗集对偶的集合对族的形式，单向 S-粗集对偶是由一串集合对构成。式(19)中，I,J 是指标集合。

由定理 3 得到：

定理 4　若元素迁移族 $\bar{F}=\phi$，则单向 S-粗集对偶与 Z. Pawlak 粗集满足：

$$\{((R,\bar{F})_o(X')_i,(R,\bar{F})^o(X')_j)\mid i\in I,j\in J\}_{\bar{F}=\phi}=(R_-(X),R^-(X)) \tag{20}$$

式(20)指出：$\bar{F}=\phi$ 的条件下，每一个单向 S-粗集对偶$((R,\bar{F})_o(X')_i,(R,\bar{F})^o(X')_j)$都回到了 Z. Pawlak 粗集$(R_-(X),R^-(X))$的“原点”。

Z. Pawlak 粗集只用一个集合对$(R_-(X),R^-(X))$表示粗集结构，一个集合对的结构来自 Z. Pawlak 粗集的“静态特性”。细心的读者可能已经发现了这个特征。

3　数据内挖掘-外挖掘

1982 年，Z. Pawlak 教授指出粗集、数据挖掘与知识发现(data mining and knowledge discovery)已成为粗集理论与应用研究中的重要研究分支之一，国内外学者在这一领域研究中取得了若干优秀成果，这些研究获得了应用。我们看一看“挖掘”一词的中文含义：“挖掘”具有“寻找”的含义。看一个例子：A 教授从实验室出来，然后回家；走到家门口，发现家门的钥匙不见了。A 教授立即想到：钥匙可能丢在实验室内，或钥匙丢在马路上(因为 A 教授没有到其它地方去)。A 教授先到实验室寻找(挖掘)，未找到；再到回家的路上找；或者在“室内”找(在实验室内找)，然后在“室外”寻找(挖掘)(在马路上找)。这个例子告诉我们：一个是内找(实验室内)，一个是外找(实验室外)。显然，一个未知的数据寻找(挖掘)的方式，应具有两种：在数据库 M 内找 $m_i\in M$，在数据库 M 外找 $m_i\overline{\in}M$。这个例子启迪我们：在数据挖掘研究中，应该进行“内挖”、“外挖”的研究；在某种意义下，“外挖”比“内挖”更重要，本文的应用部分将给出应用例子。从能查到的文献看，大多都集中在“内挖掘”讨论，而“外挖掘”的论文尚未看到，这是“数据挖掘”研究中的一种“缺失”。

从第 2 节得到：

1°　在单向 S-粗集$((R,F)_o(X^o),(R,F)^o(X^o))$中，构成$(R,F)_o(X^o),(R,F)^o(X^o)$的等价类$[x]$具有向外扩张特性，或者$[x]$内被补充了部分元素。在单向 S-粗集对偶$((R,\bar{F})_o(X'),(R,\bar{F})^o(X'))$中，构成$(R,\bar{F})_o(X'),(R,\bar{F})^o(X')$的等价类$[x]$具有向内萎缩特性，或者$[x]$内被删除了部分元素。

2°　$[x]$内被补充了部分元素；$[x]$内的元素个数增多等价于$[x]$的属性集 α 内被删除了部分属性；或者，α 内的属性个数被减少。$[x]$内被删除了部分元素；$[x]$内的元素个数减少等价于$[x]$的属性集 α 内被补充了部分属性；或者，α 内的属性个数被增多。一个例子：$\alpha^o=$红色，具有属性 α^o 的苹果 x_1,x_2,x_3 构成$[x]=\{x_1,x_2,x_3\}$；若补充属性 $\alpha'=$甜味，$\alpha''=$山东烟台，具有 $\alpha^o,\alpha',\alpha''$的苹果 x_1,x_3 构成$[x]^*=\{x_1,x_3\}$；显然，$[x]$内的元素个数减少，$[x]=\{x_1,x_2,x_3\}$变成$[x]^*=\{x_1,x_3\}$，则$[x]$的属性集 $\alpha=\{\alpha^o\}$内属性个数增多，α 变成 $\alpha^*=\{\alpha^o,\alpha',\alpha''\}$。1° 与 2° 成为本节讨论的概念依据。

在 3 节－6 节中的讨论中，Z. Pawlak 粗集$(R_-(X)=\cup[x]=\{x\mid x\in U,[x]\subseteq X\},R^-(X)=\cup[x]=\{x\mid x\in U,[x]\cap X\neq\phi\})$中的$[x]$用 m 表示，或者 $m=[x]$；单向 S-粗集$((R,F)_o(X^o)=\cup[x]=\{x\mid x\in U,[x]\subseteq X^o\},(R,F)^o(X^o)=\cup[x]=\{x\mid x\in U,[x]\cap X^o\neq\phi\})$中的$[x]$用 m^F 表示，或者 $m^F=[x]$；单向 S-粗集对偶$((R,\bar{F})_o(X')=\cup[x]=\{x\mid x\in U,[x]\subseteq X'\},(R,\bar{F})^o(X')=\cup[x]=\{x\mid x\in U,[x]\cap X'\neq\phi\})$中的$[x]$用 $m^{\bar{F}}$ 表示，或者 $m^{\bar{F}}=[x]$；不引起混乱与误解，$[x]\neq\phi$。

定义 1　称$[x]$是 D 上的一个数据 m，或者 $m=[x]$，而且

$$m=\{x_1,x_2,\cdots,x_p\} \tag{21}$$

如果 m 具有属性集 α，而且

$$\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_k\} \tag{22}$$

定义 2　称 m^F 是 m 生成的 F-数据，而且

$$m^F=\{x_1,x_2,\cdots,x_\lambda\} \tag{23}$$

如果 m^F 具有属性集 α^F，而且

$$\alpha^F=\alpha-\{\alpha_i\mid\alpha_i\in\alpha,\bar{f}(\alpha_i)=\beta_i\overline{\in}\alpha\} \tag{24}$$

式(21)，式(23)中的 p,λ 满足 $p<\lambda;\alpha^F\neq\phi$。

应当指出：$m\subset m^F$，m^F 的属性集 α^F 与 m 的属性集 α 满足 $\alpha^F\subseteq\alpha$；或者，$\mathrm{card}(\alpha^F)\leqslant\mathrm{card}(\alpha)$，card=cardinal number。

定义 3　称 $m^{\bar{F}}$ 是 m 生成的 $\bar{F}$-数据，而且

$$m^{\bar{F}}=\{x_1,x_2,\cdots,x_\gamma\} \tag{25}$$

如果 $m^{\bar{F}}$ 具有属性集 $\alpha^{\bar{F}}$，而且

$$\alpha^{\bar{F}}=\alpha\cup\{\beta_i\mid\beta_i\in V,\beta_i\overline{\in}\alpha,f(\beta_i)=\alpha'_i\in\alpha\} \tag{26}$$

式(21)，式(25)中的 γ,p 满足 $\gamma\leqslant p$，D 是有限数据论域，V 是有限属性论域。

由式(21)－式(26)与式(11)－式(15)直接得到：

定理 5($\bar{F}$-数据内挖掘定理)　给定数据 $m=\{x_1,x_2,\cdots,x_p\}\subset D,\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_k\}$是 m 的属性集，若对 α 给予部分属性补充，而且

$$\alpha^{\bar{F}}=\alpha\cup\{\beta_i\mid\beta_i\in V,\beta_i\overline{\in}\alpha,f(\beta_i)=\alpha'_i\in\alpha\} \tag{27}$$

则 $\bar{F}$-数据 $m^{\bar{F}}$ 从 m 内被挖掘，$m^{\bar{F}}\subset m$；$m^{\bar{F}}$ 具有属性集 $\alpha^{\bar{F}}$。

式中，$m^{\bar{F}}\neq\phi;p,k\in N^+$。

定理 6($\bar{F}$-数据顺序内挖掘定理)　若 $\alpha_i^{\bar{F}}$ 是 $\bar{F}$-数据 $m_i^{\bar{F}}$ 的属性集，$i=1,2,\cdots,n$，而且满足

$$\alpha_1^F \subseteq \alpha_2^F \subseteq \cdots \subseteq \alpha_{n-1}^F \subseteq \alpha_n^F \tag{28}$$

则 $\overline{F}$-数据 m_i^F 依 1,2,…,n 的顺序从 m 内被挖掘，而且

$$m_n^F \subseteq m_{n-1}^F \subseteq m_2^F \subseteq \cdots \subseteq m_1^F \tag{29}$$

显然，$\forall k \in \{1,2,\cdots,n\}$，$m_k^F \subset m$。

由式(21)－式(26)与式(1)－式(5)直接得到：

定理 7(*F*-数据外挖掘定理) 给定数据 $m=\{x_1,x_2,\cdots,x_p\} \subset D$，$\alpha=\{\alpha_1,\alpha_2,\cdots,\alpha_k\}$ 是 m 的属性集，若对 α 给予部分属性删除，而且

$$\alpha^F = \alpha - \{\alpha_i \mid \alpha_i \in \alpha, \overline{f}(\alpha_i)=\beta_i \overline{\in} \alpha\} \tag{30}$$

则 F－数据 m^F 从 m 外被挖掘，$m \subset m^F$；m^F 具有属性集 α^F。式中，$\alpha^F \neq \phi$；$p,k \in N^+$。

定理 8(*F*-数据顺序外挖掘定理) 若 α_j^F 是 F-数据 m_j^F 的属性集，$j=1,2,\cdots,n$，而且满足

$$\alpha_n^F \subseteq \alpha_{n-1}^F \subseteq \cdots \subseteq \alpha_2^F \subseteq \alpha_1^F \tag{31}$$

则 F-数据 m_j^F 依 1,2,…,n 的顺序从 m 外被挖掘，而且

$$m_1^F \subseteq m_2^F \subseteq \cdots \subseteq m_{n-1}^F \subseteq m_n^F \tag{32}$$

显然，$\forall k \in \{1,2,\cdots,n\}$，$m \subset m_k^F$。

定理 9(*F*-数据辨识定理) 若 $m_i^F, m_j^F \subset D$ 是 m 的 F-数据，α_i^F, α_j^F 分别是 m_i^F, m_j^F 的属性集，则

$$\mathrm{IDE}(m_i^F, m_j^F)_{\alpha_i^F \neq \alpha_j^F} \tag{33}$$

式中，IDE=identification，$\alpha_i^F \neq \phi$，$\alpha_j^F \neq \phi$。

定理 10($\overline{F}$-数据辨识定理) 若 $m_i^F, m_j^F \subset D$ 是 m 的 $\overline{F}$－数据，α_i^F, α_j^F 分别是 m_i^F, m_j^F 的属性集，则

$$\mathrm{IDE}(m_i^F, m_j^F)_{\alpha_i^F \neq \alpha_j^F} \tag{34}$$

式中，$m_i^F \neq \phi$，$m_j^F \neq \phi$。

由定理 5 至定理 10，得到：

定理 11(*F*-数据复制定理) 给定数据 m 的 F-数据 m_i^F，m_j^F，$i<j$；α_i^F, α_j^F 分别是 m_i^F, m_j^F 的属性集，m_j^F 是 m_i^F 的复制的充分必要条件是

$$\alpha_j^F \cup \{\alpha'_\lambda \mid \alpha_\lambda \in \alpha_i^F, \overline{f}(\alpha_\lambda)=\alpha'_\lambda \overline{\in} \alpha_i^F\} = \alpha_i^F \tag{35}$$

而且具有式(35)的 m_i^F, m_j^F 满足

$$\mathrm{UNI}(m_i^F, m_j^F) \tag{36}$$

式中，UNI=unidentification。

证明：利用第 2 节中的单向 S-粗集结构式(1)－式(5)与式(21)－式(24)得到：1° 因为 m_i^F, m_j^F，$i<j$ 是 m 的 F-数据，$m_i^F \subseteq m_j^F$；m_j^F 的属性集 α_j^F 与 m_i^F 的属性集 α_i^F 满足 $\alpha_j^F \subseteq \alpha_i^F$；或者存在属性集 $\{\alpha'_\lambda \mid \alpha_\lambda \in \alpha_i^F, \overline{f}(\alpha_\lambda) \overline{\in} \alpha_i^F\}$，把 $\{\alpha'_\lambda \mid \alpha_\lambda \in \alpha_i^F, \overline{f}(\alpha_\lambda) \overline{\in} \alpha_i^F\}$ 补充到 α_j^F 内，使得 m_j^F 是 m_i^F 的复制，或者 $m_j^F = m_i^F$；则有式(36)。2° 因为 $m_i^F \subseteq m_j^F$，若 m_i^F 的属性集 α_i^F 与 m_j^F 的属性集 α_j^F 满足式(35)，或者 m_i^F, m_j^F 具有相同的属性集，则 $m_i^F = m_j^F$，m_j^F 是 m_i^F 的复制(拷贝)。显然，若 m_j^F 是 m_i^F 的复制，则 m_j^F 与 m_i^F 满足式(36)。

定理 12($\overline{F}$-数据复制定理) 给定数据 m 的 $\overline{F}$-数据 m_i^F，m_j^F，$i<j$；α_i^F, α_j^F 分别是 m_i^F, m_j^F 的属性集，m_j^F 是 m_i^F 的复制的充分必要条件是

$$\alpha_j^F - \{\alpha'_\gamma \mid \alpha_\gamma \in \alpha_j^F, \overline{f}(\alpha_\gamma)=\alpha'_\gamma \overline{\in} \alpha_j^F\} = \alpha_i^F \tag{37}$$

而且具有式(37)的 m_i^F, m_j^F 满足

$$\mathrm{UNI}(m_i^F, m_j^F) \tag{38}$$

证明与定理 11 类似，略。

定理 11 和定理 12 指出一个事实，给出一个重要方法：

在数据库 M 内，若子数据 $m_i \in M$ 被丢失，则只要依据 m_i 具有特征集(属性集) α_i，利用 α_i 就可以找回子数据 m_i；若子数据 m_j 是进入数据库 M 的干扰数据(或入侵数据)，$m_j \in M$，则只要依据 m_j 具有特征集(属性集) α_j，利用 α_j 就可以把 m_j 从 M 内剔除(删除)，使得 $m_j \overline{\in} M$。

利用定理 5 至定理 12 得到：

命题 5 在内挖掘的 $\overline{F}$-数据 m_i^F 中，一定存在数据 m_k^F，m_k^F 的颗粒最小，或者

$$\mathrm{GRD}(m_k^F) = \min_{i=1}^{n}(\mathrm{GRD}(m_i^F)) \tag{39}$$

式中，GRD=granulation degree，$k \in \{1,2,\cdots,n\}$，$\mathrm{GRD}(m_p^F) = \mathrm{card}(m_p^F)/\mathrm{card}(m)$。

命题 6 在外挖掘的 F-数据 m_j^F 中，一定存在数据 m_λ^F，m_λ^F 的颗粒最大，或者

$$\mathrm{GRD}(m_\lambda^F) = \max_{i=1}^{n}(\mathrm{GRD}(m_i^F)) \tag{40}$$

由式(39)，式(40)得到：

内挖掘的 $\overline{F}$-数据过滤原理：具有最小颗粒的 $\overline{F}$-数据 m_k^F，从数据筛子 θ 中最先被过滤-分离。

外挖掘的 F-数据过滤原理：具有最大颗粒的 F-数据 m_λ^F，是数据筛子 θ 中被过滤-分离 m_i^F 的剩余，$i<\lambda$。

下面将第 3 节给出的讨论与结果进行抽象与理论提升。

4 数据内-外挖掘的数据圆特征

定义 4 称 y 是数据 $m=\{x_1,x_2,\cdots,x_q\}$ 特征值集合，而且

$$y=\{y_1,y_2,\cdots,y_q\} \tag{41}$$

如果 $y_i \in y$ 是 $x_i \in m$ 的特征值(x_i 的数值)，$i=1,2,\cdots,q$

式中，$y_i \in R$，R 是实数集，$i=1,2,\cdots,q$。

定义 5 称 y^F 是 F-数据 $m^F=\{x_1,x_2,\cdots,x_r\}$ 特征值集合，而且

$$y^F=\{y_1,y_2,\cdots,y_r\} \tag{42}$$

称 y^F 是 $\overline{F}$-数据 $m^F=\{x_1,x_2,\cdots,x_p\}$ 特征值集合，而且

$$y^F=\{y_1,y_2,\cdots,y_p\} \tag{43}$$

其中，式(41)－式(43)中的 p,q,r 满足 $p \leqslant q \leqslant r$；$y_k \in y^F$，$y_k \in R$；$y_t \in y^F$，$y_t \in R$。

定义 6 称 O 是数据 $m \subset D$ 生成的单位数据圆，简称 O 是数据圆；如果 O 是以坐标原点 O 为圆心，以 $\gamma = ||y||/||y|| = 1$ 为半径的圆。

式中，$||y|| = (y_1^2+y_2^2+\cdots+y_q^2)^{1/2}$ 是向量 $(y_1,y_2,\cdots,y_p)^T$ 的 2-范数，y 是特征值集合(41)生成的向量。

定义 7 称 O_F 是 F-数据 $m^F \subset D$ 生成的 F-数据圆，如果 O_F 是以坐标原点 O 为圆心，以 $\gamma^F = ||y||/||y^F||$ 为半径的圆。

式中，$||y^F|| = (y_1^2+y_2^2+\cdots+y_r^2)^{1/2}$ 是向 $y^F=(y_1,y_2,\cdots,y_r)^T$ 的 2-范数，y^F 是特征值集合式(42)生成的向量。

定义 8 称 O_F 是 $\overline{F}$-数据 $m^F \subset D$ 生成的 $\overline{F}$-数据圆，如果 O_F 是以坐标原点 O 为圆心，以 $\gamma^F = ||y||/||y^F||$ 为半径的圆。

式中，$||y^F|| = (y_1^2+y_2^2+\cdots+y_p^2)^{1/2}$ 是向 $y^F=(y_1,y_2,\cdots,y_p)^T$ 的 2-范数，y^F 是特征值集合式(43)生成的向量。

图 2 给出了数据圆 O，F-数据圆 O_F，$\overline{F}$-数据圆 O_F 的直观表示。

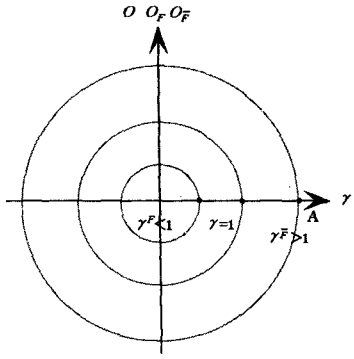

图 2 $O, O_F, O_{\bar{F}}$ 分别是数据圆、F-数据圆、$\bar{F}$-数据圆；$\gamma^F<1$ 是 O_F 的半径，$\gamma=1$ 是 O 的半径，$\gamma^{\bar{F}}>1$ 是 $O_{\bar{F}}$ 的半径；

由定义 4 至定义 8 及图 2 得到：

定理 13(数据外挖掘-内同心圆定理) 若 m_k^F 是在 m 外被挖掘的数据，则 m_k^F 生成的 F-数据圆 O_F 是 m 生成的数据圆 O 的内同心圆，而且

$$O_F \subset O \tag{44}$$

式中，借用符号“$\subset$”表示 O_F 被嵌套在 O 内。

定理 14(数据内挖掘-外同心圆定理) 若 m_k^F 是在 m 内被挖掘的数据，则 m_k^F 生成的 $\bar{F}$-数据圆 $O_{\bar{F}}$ 是 m 生成的数据圆 O 的外同心圆，而且

$$O \subset O_{\bar{F}} \tag{45}$$

式中，借用符号“$\subset$”表示 $O_{\bar{F}}$ 被嵌套在 O 外。

定理 13 及定理 14 的证明由图 2 直观地得到，证明略。

对定理 13 及定理 14 再进行讨论，得到：

定理 15(数据内挖掘-单位离散区间外点定理) 设 $(0,1]$ 是数值 0 与数据圆 O 的半径 $\gamma=1$ 构成的单位离散区间，若 m^* 是在 m 内被挖掘的数据，则

1° m^* 生成的数据圆 O^* 的半径 γ^* 是单位离散区间 $(0,1]$ 的外点，或者

$$\gamma^* \overline{\in} (0,1] \tag{46}$$

2° m^* 是 m 的 $\bar{F}$-数据，而且

$$m^* = m^{\bar{F}} \tag{47}$$

证明：因为 m^* 是在 m 内被挖掘的数据，$m^* \subseteq m$；$(0,1]$ 是数值 0 与定义 6 中的 $\gamma=||y||/||y||=1$ 构成的单位离散区间。设 $y^*=\{y_1, y_2, \cdots, y_p\}$ 是 m^* 的特征值集合，$y=\{y_1, y_2, \cdots, y_q\}$ 是 m 的特征值集合，由定义 8 得到：$\gamma^*=||y||/||y^*||>1$，则有 $\gamma^* \overline{\in} (0,1]$。因为 $m^* \subseteq m$，由式(25)，式(26)得到：$m^*=m^{\bar{F}}$。

定理 16(数据外挖掘-单位离散区间内点定理) 设 $(0,1]$ 是数值 0 与数据圆 O 的半径 $\gamma=1$ 构成的单位离散区间，若 m^o 是在 m 外被挖掘的数据，则

1° m^o 生成的数据圆 O^o 的半径 γ^o 是单位离散区间 $(0,1]$ 的内点，或者

$$\gamma^o \in (0,1] \tag{48}$$

2° m^o 是 m 的 F-数据，而且

$$m^o = m^F \tag{49}$$

证明：与定理 15 类似，证明略。

定理 17(内点-外点重合定理) 设 $m^F, m^{\bar{F}}$ 分别是在 m 内挖掘的数据，在 m 外挖掘的数据，若

$$\mathrm{UNI}(m^F, m^{\bar{F}}) \tag{50}$$

则

$$\gamma^F = \gamma^{\bar{F}} \tag{51}$$

式中，$\gamma^F \overline{\in} (0,1]$，$\gamma^{\bar{F}} \in (0,1]$。

定理 18(数据圆重合定理) 设 $m^F, m^{\bar{F}}$ 分别是在 m 内挖掘的数据，在 m 外挖掘的数据，若

$$\mathrm{UNI}(m^F, m^{\bar{F}}) \tag{52}$$

则

$$O_F = O = O_{\bar{F}} \tag{53}$$

式中，$O, O_F, O_{\bar{F}}$ 分别是数据圆、F-数据圆、$\bar{F}$-数据圆。

定理 17 和定理 18 的证明是直接、容易的，证明略。

由定义 4 至定义 8 及定理 13 至定理 18 的得到：

数据挖掘的数据圆准则

若 O^* 是数据 m^* 生成的数据圆，而且满足

$$O^* \subset O \text{ 或 } O \subset O^* \tag{54}$$

则 m^* 在 m 外部或者 m^* 在 m 内部；m^* 的属性集 α^* 与 m 的属性集 α 具有关系 $\alpha^* \cap \alpha \neq \phi$。

5 数据内-外挖掘与数据圆应用

为了简单，又不失一般性，在本节的讨论中，只给出数据内挖掘与数据圆的简单应用，应用例子中属性 $\alpha_i \in \alpha$，元素 $x_j \in X$ 的名称，略。

2009 年，重庆市进行震惊全国的“打黑除恶行动”，一些“横行乡里、鱼肉百姓”的犯罪分子被绳之以法，维护了法律尊严，偿还给市民们和谐生存的生活空间，受到全国人民的瞩目。

w 是这次“打黑除恶”行动中的一个案件，$x_1, x_2, x_3, x_4, x_5, x_6, x_7$ 是 w 的犯罪嫌疑人，他们具有犯罪事实(属性)$\alpha_1, \alpha_2, \alpha_3, \alpha_4$，如表 1 中所列

表 1 案件 w 的犯罪嫌疑人 x_i 与犯罪事实(属性)α_i

案件 w							
[x]	x_1	x_2	x_3	x_4	x_5	x_6	x_7
α	α_1	α_2	α_3	α_4			

显然，$x_1, x_2, x_3, x_4, x_5, x_6, x_7$ 关于 $\alpha=\{\alpha_1, \alpha_2, \alpha_3, \alpha_4\}$ 构成等价类(犯罪团伙)$[x]=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$。$[x]$ 中 x_1—x_7 的犯罪有“轻”、“重”之分；因此，需要找出 $[x]$ 中的“首犯”与“从犯”。对于“首犯”按法律量刑，对于“从犯”进行“教育、释放”。刑警人员为此又进行了大量调查、取证；获得了新的证据(属性)$\beta_1, \beta_2, \beta_3$；表 1 变成表 2。

表 2 案件 w 的犯罪事实补充与案件首犯发现

案件 w							
$[x]^*$	x_1	x_4					
α^*	α_1	α_2	α_3	α_4	β_1	β_2	β_3

显然，对犯罪事实集(犯罪证据集，属性集 α)，给予属性补充，或者 $\alpha^*=\alpha \cup \{\beta_i \mid \beta_i \in V, \beta_i \overline{\in} \alpha, f(\beta_i)=\alpha'_i \in \alpha\}$，则“首犯”$[x]^*$ 从 $[x]$ 内被挖掘；或者数据 $m^*=[x]^*$ 从数据 $m=[x]$ 内被挖掘出来。我们回到本文第 2 节中，$[x]=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$ 变成 $[x]^*=\{x_1, x_4\}$ 是单向 S-粗集对偶具有的特征：元素 x_2, x_3, x_5, x_6, x_7 被元素迁移 $\bar{f} \in \bar{F}$ 从 $[x]$ 内迁出到 $[x]$ 外，或者 $X'=X-\{x \mid x \in X, \bar{f}(x)=u \overline{\in} X\}$。利用第 3 节中的内挖掘的 $\bar{F}$-数据过滤原理得到：$[x]^*$ 从 $[x]$ 内被过滤-分离，或者利用式(39)，$m^*=[x]^*$ 具有颗粒度 $\mathrm{GRD}(m^*=[x]^*)=2/7=0.28<1$；显然，$m^*$ 构成的数据圆 O^* 是 m 构成的数据圆 O 的外同心圆。

应当指出：案件 w 的原始犯罪嫌疑人并不只是 x_1-x_7，应当比 x_1-x_7 还要多，如何找到 w 的原始犯罪嫌疑人，或者利用 $[x]$ 外挖掘找到 $[x]'$，$[x]\subset[x]'$，$m\subset m'$；把上面的讨论反过就可得到，这些讨论留给读者。

本文的应用讨论很容易移植到动态数据系统（动态信息系统）的识别研究、动态数据库的子数据库的识别研究，请读者试一试，很简单。

6 Z. Pawlak 粗集的精华与本质

杰出的学者 Z. Pawlak 教授的原创性工作：提出粗集（Rough sets），给出粗集的结构与应用，使人们看到集合论中还有这样的新鲜概念，从而引起学术界对 Z. Pawlak 教授的尊敬。这启迪着人们在粗集这个“科学的百花园”内去耕耘。我们自然想到 Z. Pawlak 教授提出的粗集其“精华与本质”是什么？Z. Pawlak 教授做了一件什么事？我们用下面的 2 个简图来说明。

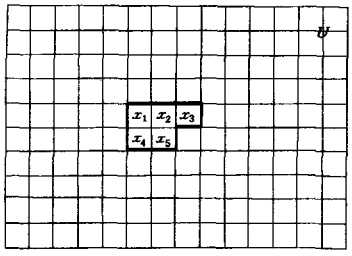

图 3　x_1,x_2,x_3,x_4,x_5 构成普通集合 $X=\{x_1,x_2,x_3,x_4,x_5\}$

图 3 定义成一张牛皮，图中的每一个小方块定义成能做 24 码的一双皮鞋；显然，粗黑线表示的牛皮，做而只能做 5 双皮鞋 x_1,x_2,x_3,x_4,x_5，不多不少。显然，x_1,x_2,x_3,x_4,x_5 构成了普通集合 $X=\{x_1,x_2,x_3,x_4,x_5\}$（$X$ 是数学分析，离散数学，高等数学中的集合），集合 X 很精确；天底下有这么多精确的事吗？哪一头牛杀掉之后，牛皮是图 3 中表示的方方正正的牛皮？我们再看图 4。

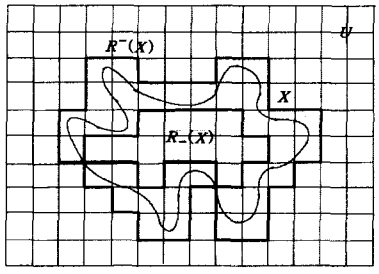

图 4　边界不规则集合 X 与元素分布；下近似 $R_-(X)$，上近似 $R^-(X)$ 的直观表示

图 4 给出一张自然牛皮 X，这张牛皮能做多少双 24 码的皮鞋？显然，我们不能像图 3 那样给出精确回答，更不能用数学分析、离散数学、高等代数中的集合概念给出回答。在 1982 年之前，国内外经典数学界对此问题回避，或者对此问题视而不见；更没有人对此近似问题给出讨论，这不能不说是数学研究中的一种“缺失”。1982 年 Z. Pawlark 教授巧妙地定义了 $X\subset U$ 的下近似 $R_-(X)=\bigcup[x]=\{x\mid x\in U,[x]\subseteq X\}$，$X\subset U$ 的上近似 $R^-(X)=\bigcup[x]=\{x\mid x\in U,[x]\cap X\neq\phi\}$；用 $R_-(X)$，$R^-(X)$ 共同“近似逼近”集合 X。集合 X 在图 4 中用细实线表示，下近似 $R_-(X)$ 在 X 内用粗实线表示，上近似 $R^-(X)$ 在 X 外用粗实线表示。显然，若下近似 $R_-(X)$ 向外扩张，扩张到 X；上近似 $R^-(X)$ 向内收缩，收缩到 X，则有 $R_-(X)=X=R^-(X)$；粗集回到经典集的“原点”。因此，“近似，逼近”是 Z. Pawlak 粗集的“精华与本质”。如果 $R_-(X)=R^-(X)=X$，则 Z. Pawlak 粗集就是普通集合 X。Z. Pawlak 粗集中的“近似，逼近”思想使我们想到了定积分：

$$I=\int_a^b f(x)\mathrm{d}x \tag{55}$$

式(55)表示一个曲边梯形的面积；显然，用式(55)去求一个曲边梯形面积时，“近似，逼近”的思想成为解决问题的关键。

解读 Z. Pawlak 教授的工作，认识 Z. Pawlak 教授这一杰出的学术贡献与原创性的学术思想，引起活着的人们对这位已经谢世长者的怀念与敬仰。Z. Pawlak 教授的原创性工作为后人的科学研究搭建了一个平台，平台中的概念启迪着后来者。

7 讨论与建议

Z. Pawlak 教授提出粗集，这一成功的研究开辟了数学的理论研究与应用研究的新分支，给计算机科学、信息科学、系统科学研究添加了一个新的数学工具。Z. Pawlak 教授的工作是否是很完美了？人们自然会想到这个问题，这是因为人类的认识永远不会停留在同一个水平上。2002 年，作者把“动态特性”引入到 Z. Pawlak 粗集中，改进了 Z. Pawlak 粗集提出 S-粗集（单向 S-粗集、单向 S-粗集对偶、双向 S-粗集），给出它们在计算机科学、信息科学、系统科学中的多个应用[21,22]；作者的这些工作是在 Z. Pawlak 工作引领之下得到的。单向 S-粗集、单向 S-粗集对偶、双向 S-粗集扩展了 Z. Pawlak 粗集的应用空间；S-粗集已形成了 Z. Pawlak 粗集理论与应用的新分支。我们面对的信息系统具有动态特性，S-粗集已成为动态信息系统研究的一个新的数学工具与方法。在 S-粗集理论与应用研究中，下列问题可能成为新的创新点。

动态信息的依赖识别

如果把推理模式引入到 S-粗集中，建立动态推理过程（以元素迁入，元素迁出为依据），有可能得到“信息依赖识别”的一系列优秀结果。

系统状态粗辨识

在 t 时刻，系统输出状态用数据 $[x]=\{x_1,x_2,\cdots,x_k\}$ 表示（x_i 具有数值 y_i），因为系统内部的系统参数发生了变化，$[x]$ 变成 $[x]^F=\{x_1,x_2,\cdots,x_r\}$，$r<k$；或者 $[x]$ 变成 $[x]^F=\{x_1,x_2,\cdots,x_\lambda\}$，$k<\lambda$；这个现象被经常做系统的学者们遇到。$[x]$ 变成 $[x]^F$，$[x]^F\subset[x]$；$[x]$ 变成 $[x]^F$，$[x]\subset[x]^F$ 正是单向 S-粗集对偶、单向 S-粗集的特征；如果把单向 S-粗集对偶、单向 S-粗集引入到这个现象中，对这个现象给出讨论，能够得到一些耳目一新的结果。

粗可能性与它的度量

在可能性理论中，如果给出这样的定义：$x_1,x_2,\cdots,x_k$ 个人构成的集合 $X=\{x_1,x_2,\cdots,x_k\}$ 在给定的条件 η 下完成工程 w 的可能性是 ρ，$\rho\in(0,1)$。若 X 内有 q 个人生病，X 变成 $X^F=\{x_1,x_2,\cdots,x_{k-q}\}$，

集合 $X^F=\{x_1,x_2,\cdots,x_{k-q}\}$ 在给定的条件 η 下完成工程 w 的可能性还是 ρ，$\rho\in(0,1)$ 吗？反之，为解决就业问题，X 内增加 t 个人，X 变成 $X^F=\{x_1,x_2,\cdots,x_k,x_{k+1},x_t\}$，$X^F=\{x_1,x_2,\cdots,x_k,x_{k+1},x_t\}$ 在给定条件 η 下完成工程 w 的可能性还是 ρ，$\rho\in(0,1)$ 吗？显然不是。这种具有动态特征的可能性，能到处遇到。如果把单向 S-粗集对偶、单向 S-粗集交叉到可能性研究中，能够得到粗可能性的概念与应用。事实上数对 $(\rho_F,\rho_{\overline{F}})$ 就构成了粗可能性，它是一种具有数对形式的度量。

数据变换与它的伪装

在单向 S-粗集中，数据 $m=[x]$ 发生补充元素的变化，或者 $[x]$ 外一些新数据通过元素迁移 $f\in F$ 的作用进入到 $[x]$ 内，$[x]$ 变成 $[x]^F=m^F$；显然，m^F 可以看作 m 的 F-伪装；在单向 S-粗集对偶中，数据 $m=[x]$ 发生删除元素的变化，或者 $[x]$ 内的一些数据通过元素迁移 $\overline{f}\in\overline{F}$ 作用离开 $[x]$，$[x]$ 变成 $[x]^F=m^F$；显然，m^F 可以看作 m 的 $\overline{F}$-伪装。因此，把数据 m 伪装成数据 (m^F,m^F)，在计算机网络中传递 (m^F,m^F)，达到“以假乱真”的目的，真实的数据 m 就潜藏在 (m^F,m^F) 内，并获得了保护；从 (m^F,m^F) 中获取真实的数据 m 是困难的。这里仅给出 4 个待讨论的问题，这些问题均具有很好的理论与应用前景；从这 4 个问题中将得到“耳目一新的原创性”的成果，建议年轻的学者们不妨作些尝试，顺着这条路走一走，能够得到一些令人高兴的结果。

8 对 Z. Pawlak 粗集、S-粗集再扩展

面对 Z. Pawlak 粗集、S-粗集，我们提出这样的问题：Z. Pawlak 粗集、S-粗集能用来发现信息系统中的未知信息规律吗？显然不能，这是因为 Z. Pawlak 粗集、S-粗集不具有规律特征（S-粗集比 Z. Pawlak 粗集只多了一个动态特性）。什么是规律？以系统科学的观点，$[a,b]$ 区间上的函数（连续函数，离散函数）是 $[a,b]$ 区间上的一个规律（连续规律，离散规律）。例如，在初中物理课上，初速度 v_0 不为零的质点加速运动，可用 $S=v_0t+\frac{1}{2}at^2$ 来表示，二次函数 $v_0t+\frac{1}{2}at^2$ 表达了质点的运动规律。显然，应当把 Z. Pawlak 粗集、S-粗集再改进，使得粗集能应用于寻找未知信息规律研究；在某种意义上说，规律挖掘比数据挖掘更重要。2005 年，作者把函数概念引入到 S-粗集、Z. Pawlak 粗集中，提出了函数 S-粗集[11-16,18-22] 函数 S-粗集是改进 S-粗集得到的；提出的函数粗集[11,21,22] 是改进 Z. Pawlak 粗集得到的，文献[21,22]给出函数 S-粗集在多个领域中的应用，读者能够看到：函数 S-粗集比 S-粗集、Z. Pawlak 粗集具有更大的应用空间。例如，函数 S-粗集能应用到信息图像隐藏、信息图像伪装、信息图像变换的多个具有应用价值的领域。关于函数 S-粗集的结构、特征、应用，将在后面讨论。

参 考 文 献

[1] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 32(11): 341-356

[2] Shi Kaiquan. S-rough sets and its applications in diagnosis-recognition for disease[J]. IEEE Proceedings of the First International Conference on Machine Learning and Cybernetics, 2002, 1(1): 50-54

[3] 史开泉，崔玉泉. S-粗集与它的结构[J]. 山东大学学报：理学版，2002, 37(6): 471-474

[4] Shi Kaiquan, Cui Yuquan. F-decomposition and $\overline{F}$-reduction of S-rough sets[J]. An International Journal Advances in Systems Science and Applications, 2004, 4(4): 487-499

[5] Shi Kaiquan, Chang Tingcheng. One direction S-roughsets[J]. International Journal of Fuzzy Mathematics, 2005, 13(2): 319-334

[6] Shi Kaiquan. Two direction S-rough sets[J]. International Journal of Fuzzy Mathematics, 2005, 13(2): 335-349

[7] 史开泉，崔玉泉. S-粗集与它的分解-还原[J]. 系统工程与电子技术，2005, 27(4), 644-651

[8] Shi Kaiquan. S-rough sets and knowledge separation[J]. Journal of Systems Engineering and Electronics, 2005, 16(2): 237-240

[9] Hu Haiqing, Wang Hongyu, Shi Kaiquan. Two direction S-rough recognition of knowledge and recognition model[J]. An International Journal Advances in Systems Science and Applications, 2005, 5(3): 368-374

[10] 史开泉. S-粗集与新材料发现-识别[J]. 系统工程与电子技术，2006, 28(3): 383-388

[11] Shi Kaiquan. Function S-rough sets and function transfer[J]. An International Journal Advances in Systems Science and Applications, 2005, 5(1): 1-8

[12] 张萍，史开泉，卢昌荆. 函数 S-粗集与粗规律挖掘-分离[J]. 系统工程与电子技术，2005, 27(11): 1899-1902

[13] 史开泉，姚炳学. 函数 S-粗集与规律辨识[J]. 中国科学 E：信息科学，2008, 38(4): 553-564

[14] Shi Kaiquan, Yao Bingxue. Function S-rough sets andlaw identification[J]. Science in China F: Information Sciences, 2008, 51(5): 499-510

[15] 史开泉，赵建立. 函数 S-粗集与隐藏规律安全-认证[J]. 中国科学 E：信息科学，2008, 38(8): 1234-1243

[16] Shi Kaiquan, Zhao Jianli. Function S-rough sets and security-authentication of hiding law[J]. Science in China F: Information Sciences, 2008, 51(7): 924-935

[17] Zhang Ling, Shi Kaiquan. Security transmission and recognition of F-konwledge[J]. Journal of Systems Engineering and Electronics, 2009, 20(4): 877-882

[18] Qiu Jinming, Shi Kaiquan. F-rough law and the disovery of rough law[J]. Journal of Systems Engineering and Electronics, 2009, 20(1): 81-89

[19] Zhou Houyong, Huang Shunliang, Shi Kaiquan. Hiding dependence-discovery of F-hiding laws System law[J]. Journal of Systems Engineering and Electronics, 2009(3): 543-550

[20] Li Dongya, Ren Xuefang, Shi Kaiquan. Rough law generation and its separation-recognition[J]. Journal of Systems Endineering and Electronics, 2009, 20(6): 1239-1246

[21] 史开泉，崔玉泉. S-粗集与粗决策[M]. 北京：科学出版社，2006: 155-165

[22] 史开泉，姚炳学. 函数 S-粗集与系统规律挖掘[J]. 科学出版社，2007: 147-198