# 基于Rough Set和neural network组合数据挖掘

王志明

（怀化职业技术学院，湖南 怀化 418000）

**摘 要**：提出了一种基于rough set和neural network的数据挖掘新方法。首先利用粗集理论对原始数据进行一致性属性约简，然后使用神经网络对数据进行学习，并同时完成属性的不一致约简，最后再由粗集对神经网络中的知识进行规则抽取。该方法充分融合了粗集理论强大的属性约简、规则生成能力和神经网络优良的分类、容错能力。实验表明，该方法快速有效，生成规则简单准确，具有良好的鲁棒性。

**关键词**：数据挖掘；粗集理论；神经网络；分类

**中图分类号**：TP391.03 **文献标识码**：A **文章编号**：1673-9833(2007)02-0079-05

## An Approach of Data Mining Based on Rough Set and Neural Network

Wang Zhiming

（Huaihua Vocational and Technical College，Huaihua Hunan 418000，China）

**Abstract**：A new method of data mining based on rough set and neural network is proposed. Based on the rough set theory, attribute reduction is processed on data under the consistent conditions. Then neural network is used to study and predict data, as well as reduce the attributes under the inconsistent conditions. Finally rule knowledge in the neural network is extracted by using rough set theory. The method mixes rough set's strong attribute reduction, rule extraction ability and neural networks classification, robustness ability. Experimental results show that this algorithm can produce more effective and simpler rules quickly and possess good robustness.

**Key words**: data mining; rough set; neural network; classification

## 0 引言

数据挖掘（Data Mining，简称DM）也称数据库知识发现（Knowledge Discovery in Database，简称KDD），是一个从数据库的数据中提取隐含的有用信息（或知识）的过程。面对高维的、超大规模的数据库，如何建立有效的、可扩展的数据挖掘算法是数据挖掘研究的方向之一。

粗集（Rough Set，简称RS）理论是20世纪80年代初由Pawlak提出的一种处理模糊性和不确定性的数学工具。在处理大数据量、消除冗余信息等方面，粗集理论有着良好的效果，因而被广泛应用于数据挖掘的数据预处理、数据缩减、规则生成、数据依赖关系发现等方面。但是，由于粗集理论对错误描述的确定性机制过于简单，而且在约简的过程中缺乏交互验证功能，因此当数据中存在噪声时，其结果往往不稳定，精度不高。神经网络由于分类精度高，鲁棒性强等优点，在机器学习、模式识别等领域得到了广泛的应用。然而，神经网络面对数据挖掘中的高维和超大规模问题，其学习速度缓慢，规则生成困难等缺陷表现得更为明显，原有的神经网络算法在效率和可扩展方面都会出现问题。

由于粗集理论与神经网络具有很强的优势互补性，因此两种技术的有效结合是当前的一个研究热点。使用粗集理论对输入到神经网络的数据进行属性约简和属性域约简，使得网络的学习速度大大加快，分类精度显著提高，但文章没有涉及最终知识的获取问题。针对以上问题，本文提出一种融合粗集理论和

**收稿日期**：2007-01-27

**作者简介**：王志明（1970-），男，湖南双峰人，怀化职业技术学院副教授，硕士研究生，主要研究方向为智能信息处理.

神经网络的数据挖掘新方法，应用于从大型数据库中挖掘分类规则。其主要思想是首先由粗集理论对数据库进行初步约简；然后借助于神经网络在自学习的过程中完成对数据库的进一步属性约简，并过滤数据中的噪声数据，最后由粗集理论对约简后的数据库进行规则抽取，得到最终的挖掘知识。通过与现有的数据挖掘方法的实验比较，验证了本文方法的有效性。

# 1 基本原理

## 1.1 决策表

数据库可以分为两类：一类是一致性的，另一类是不一致性的。如果一个数据库中存在不同的实例，他们具有相同的条件属性值而具有不同的分类，则这类数据库是不一致性的，否则为一致性的。

分类问题中数据库大多可以用决策表的形式给出。令 $T=(U, A)$是一个决策表，$U=\{x_1, x_2,\cdots,x_n\}$是实例的集合，也称为论域。$A = C \cup D$ 是属性集合，$C=\{c_1,c_2,\cdots,c_N\}$是条件属性的集合，$D=\{d_1,d_2,\cdots,d_M\}$是决策属性的集合。分类问题中 $D$ 的每一种不同取值被归为一类，则 $D$ 可以归并为一个单元素集 $D=\{class\}$。一般地，表中的列表示属性，行表示实例，表中的每个值都是对应行（实例）在对应列（属性）下的值，也称属性值。如决策表 1 所示，$a(x_2)=1$ 表示了实例 $x_2$ 在属性 $a$ 上的属性值为 1。

**表 1 一致性决策表**

**Table 1 Consistency decision table**

| $U$ | $a$ | $b$ | $c$ | $class$ |
|---|---|---|---|---|
| $x_1$ | 1 | 0 | 1 | 0 |
| $x_2$ | 1 | 1 | 0 | 0 |
| $x_3$ | 0 | 1 | 1 | 1 |

由以上定义可知，表 1 是一致性决策表，表 2 是不一致性决策表。因为表 2 中的实例 $x_1,x_3$ 在条件属性 $a$，$b$，$c$ 上的属性值均为 1，0，1，而其决策属性却不相同，$class(x_1)=0$，$class(x_3)=1$。

**表 2 不一致性决策表**

**Table 2 Inconsistency decision table**

| $U$ | $a$ | $b$ | $c$ | $class$ |
|---|---|---|---|---|
| $x_1$ | 1 | 0 | 1 | 0 |
| $x_2$ | 1 | 1 | 0 | 0 |
| $x_3$ | 1 | 0 | 1 | 1 |

如果没有特别说明，那么本文设定的初始数据库是一致性的。

## 1.2 方法与主要步骤

本文基于粗集理论（RS）和神经网络（NN）的数据挖掘方法（RNDM）包括了 3 个主要阶段：

1) 使用粗集理论对初始决策表 $T$ 进行一致性属性约简，得到决策表 $T_1$。一致性属性约简是指在属性的约简过程中保持决策表的一致性。

2) 使用神经网络对决策表 $T_1$ 进行不一致属性约简及实例约简，得到决策表 $T_2$。首先使用神经网络对决策表中的数据进行学习；然后对作为网络输入的属性进行逐步删除，直到其分类精度明显下降时停止，得到网络 $N$，设网络 $N$ 的属性约简子集为 $C$，不能被网络 $N$ 正确分类的实例集为 $E$；最后在决策表中只保留 $C$ 中含有的属性的列，并删除 $E$ 中含有的实例的行，得到进一步约简的决策表 $T_2$。

3) 使用粗集理论从决策表 $T_2$ 中抽取规则，得到最终的挖掘知识—规则集方法的主要步骤，如图 1 所示。

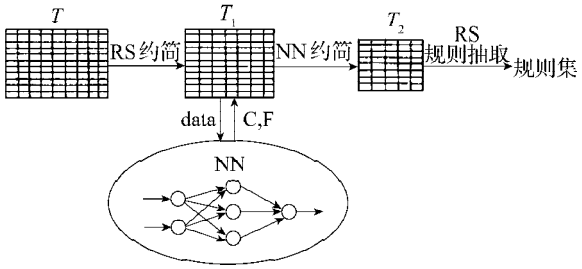


**图 1 RNDM 方法的主要步骤**

**Fig. 1 Major steps of RNDM methods**

## 1.3 属性约简

目前的基于粗集理论的属性约简方法大多是在保持数据库一致性的前提下进行的。使用粗集理论进行不一致属性约简时，由于其定义的分类边界过于简单，对错误的描述机制薄弱，而且在约简过程中缺乏有效的交互验证方法，因而会出现约简结果不稳定、分类精度不高的问题。而基于神经网络的属性约简(也称特征选择) 则因为数据挖掘问题中属性众多、数据规模庞大，而不可避免地存在网络结构庞大、约简速度极为缓慢等问题。

本文有效融合粗集理论和神经网络两种技术的优势，提出分两步进行属性约简。首先由粗集理论完成对属性的一致性约简，并由粗集理论对约简后的属性子集中的属性进行重要性排序。然后由神经网络按照重要性由小到大的顺序对属性进行不一致属性约简。借助于粗集理论对数据的预处理以及提供的先验信息(重要性排序)，大大加快了神经网络学习的速度；同时利用神经网络强大的分类，容错能力有效保证了不一致属性约简的正确性和稳定性。

## 1.4 规则抽取

神经网络的知识分布于其网络结构和定向连接的权重中，因此从训练后的网络中获取知识是十分困难的。文献[5]中提出了一种从训练后的 3 层前馈网络中抽取分类规则的方法。首先对每个隐层结点的输出值进行离散化；然后分别导出隐层结点与输出层结点之间的规则（规则集 1）和输入层结点与隐层结点之间的规则（规则集 2）；最后将两部分规则进行合并得到

最终的分类规则。这其中隐结点输出值的离散化过程不仅繁琐，而且会丢失信息。另外，当网络规模较大时，规则集 1 和规则集 2 的规模将是巨大而不实用的。

使用粗集理论从决策表中抽取规则相对要简单和直接。从图 1 可以看出，神经网络(NN) 可以看作是对决策表 $T_1$ 中知识的一种学习。对于不能分类的实例集 $E$，其知识不能被 NN 学习到。因此能被 NN 正确分类的实例均包含在 $T_1$ 约简后得到的决策表 $T_2$ 中，所以 NN 可以看作是 $T_2$ 中的知识的一种映射。本文使用粗集理论从 $T_2$ 中抽取的规则集知识则是对 $T_2$ 中的知识的另一种映射，它避开了从神经网络中抽取规则的困难，但利用了神经网络的容错能力对决策表中的噪声进行了过滤，在保证获得规则简单准确的同时，显著提高了抽取规则的速度。

# 2　基于 rough set 和 neural network 数据挖掘算法

## 2.1　基于粗集理论和神经网络的数据挖掘主算法

步骤 1：使用基于粗集理论的属性约简子算法对初始决策表 $T$ 进行约简，得到决策表 $T_1$ ，其条件属性子集 $C=\{c_1, c_2,\cdots, c_K\}$，满足 $p(c_1)\leqslant p(c_2)\leqslant\cdots\leqslant p(c_K)$。$p(c_i)$为属性 $c_i$ 的分类重要性函数。

步骤 2：将决策表 $T_1$ 中的实例分为两个集合，即训练集 $S_1$ 和交互验证集 $S_2$。令 $\Delta A$ 为 $S_2$ 上分类精确度的最大允许下降。

步骤 3：以决策表 $T_1$ 的条件属性作为神经网络的输入结点，决策属性作为输出结点，构造 3 层前馈神经网络 $N$。使用 BP 算法训练 $N$ 使其在 $S_1$ 上达到要求的最小分类准确度。令 $A_2$ 为网络 $N$ 在 $S_2$ 上的分类准确度。选择 $T_1$ 条件属性子集的第一个属性为 $C_{\text{choose}}$。

步骤 4：a）删除网络 $N$ 中的 $C_{\text{choose}}$ 属性对应的输入结点及与其相连的所有连接，得到网络 $N_1$。b) 重新训练 $N_1$ 使其在 $S_1$ 上达到要求的分类准确度。令 $A_{2'}$ 为网络 $N_1$ 在 $S_2$ 上的分类准确度，$DEC=(A_2-A_{2'})A_2$。c) 如果 $DEC\leqslant\Delta A$，则 $N\leftarrow N_1$；$A_2\leftarrow\max\{A_2, A_{2'}\}$；$C\leftarrow C-\{C_{\text{choose}}\}$。选择 $C$ 中的第一个属性为 $C_{\text{choose}}$，返回步骤 a)；否则，进行下一步。d) 如果 $C_{\text{choose}}$ 不是 $C$ 中最后一个属性，则选择 $C$ 中 $C_{\text{choose}}$ 的下一个属性为 $C_{\text{choose}}$，返回步骤 a)；否则，进行下一步。

步骤 5：设网络 $N$ 不能正确分类的实例集为 $E$。在决策表中只保留 $C$ 含有的属性的列，合并重复的实例行，删除 $E$ 中包含的实例的行，得到决策表 $T_2$。

步骤 6：使用基于粗集理论的规则抽取子算法对 $T_2$ 进行规则抽取，得到最终的规则知识。

需要指出的是，步骤 5 中将网络 $N$ 不能正确分类的实例从决策表 $T_1$ 中删除，不仅有效地过滤了 $T_1$ 中包含的噪声实例，而且保证了决策表 $T_2$ 的一致性。因为 $T_2$ 中的实例均可以被网络 $N$ 正确分类，所以它肯定不包含条件属性值完全相同而决策属性值不同的两个不同实例。

## 2.2　基于粗集理论的属性约简子算法

Sknowron 提出的差别矩阵为属性约简提供了很好的思路。令决策表 $T=(U, C, D)$，实例集合 $U=\{x_1, x_2, \cdots, x_n\}$，条件属性 $C=\{c_1, c_2, \cdots, c_n\}$，$D$ 为决策集合，差异矩阵 $\boldsymbol{M}$ 定义为：

$$M(i,j)=\begin{cases}\{c_k\in C: c_k(x_i)\neq c_k(x_j)\}, & D(x_i)\neq D(x_j);\\ 0, & D(x_i)=D(x_j);\\ -1, & \left(C(x_i)=C\left(x_j\right)\right)\cap\left(D\left(x_i\right)\neq D\left(x\left(x_j\right)\right)\right)。\end{cases}$$

其中 $i$，$j$=1，2，…，$n$；$k$=1，2，…，$N$。

借助于差异矩阵，本文给出基于粗集理论的属性约简子算法的具体步骤：

步骤 1：由决策表 $T$ 构造差异矩阵 $\boldsymbol{M}$，初始化属性约简子集 $R=\Phi$。

步骤 2：令 $p(c_k)$ 为属性 $c_k$ 在 $\boldsymbol{M}$ 中出现的次数。计算出现在 $\boldsymbol{M}$ 中的所有 $c_k$ 的 $p(c_k)$，得到 $p(c_k)$中最大值所对应的属性 $c_q$。如果 $c_q$ 只有一个，将属性 $c_q$ 选为 *SELECT*；如果有多个，则从中随机选取一个属性作为 *SELECT*。

步骤 3：$R\leftarrow R\cup\{SELECT\}$。令 $\boldsymbol{M}$ 中所有含有 *SELECT* 的项 $M(i, j)=\Phi$。

步骤 4：如果 $\boldsymbol{M}$ 中仍有含有属性的项，转到步骤 2；否则，进行下一步。

步骤 5：$R$ 为决策表 $T$ 的一个属性约简子集。只保留 $T$ 中 $R$ 包含的属性的列，删除重复的实例，将新的决策表赋给 $T_1$ 。

一般地，基于粗集理论的属性约简算法总是将条件集合中的核属性(即那些唯一能够将两个或多个实例区分开来的属性) 先选入约简子集，这对保持决策表的一致性是必需的。但是由于噪声等因素的存在，在进行不一致属性约简时，某些核属性常常会被约简掉，这说明其对整个决策表的分类作用不大。步骤 2 中以属性在差异矩阵中出现的次数多少表示该属性对决策表整体分类的贡献大小。因此，越早被选中成为 $R$ 中元素的属性，其分类重要性越大。实践证明，本文算法更加符合实际数据库，可大大提高以后的不一致属性约简表。

## 2.3　基于粗集理论的规则抽取子算法

粗集理论中，从决策表中抽取规则的过程实质上是一个对决策表进行值约简的过程。本文采用文献[8]中的值约简算法来抽取规则。其主要步骤为：

步骤 1：对决策表中的条件属性进行逐列考察。如果删除该属性列后，若产生冲突实例，则保留冲突实例的原该属性值；若为产生冲突但含有重复实例，则

将重复实例的该属性值标为“3”；对其它实例，将该属性值标为“?”。

步骤2：删除可能产生的重复实例，并考察每条标记“?”的实例。若仅由未被标记的属性值即可判断出决策，则将“?”标记为“3”，否则，修改为原属性值；若某个实例的所有条件属性均被标记，则将标有“?”的属性项修改为原属性值。

步骤3：删除所有条件属性均被标为“3”的实例及可能产生的重复实例。

步骤4：如果两个实例仅有一个条件属性值不同，且其中一个实例该属性被标为“3”，那么，对该实例如果可由未被标记的属性值判断出决策，则删除另外一个实例；否则删除本实例。

经过值约简后的决策表，每个实例代表一条规则，每个实例中没有被标记为“3”的属性个数组为该条规则的条件数。如表3为一个值约简后的决策表。其中有1、2、3三个实例，$c_1$、$c_2$、$c_3$为条件属性，*class*为决策属性。则抽取的规则为：

规则1：If ($c_1$ = 1) then *class* = 0；

规则2：If ($c_2$ = 0) ∧ ($c_3$ = 0) then *class* = 0；

规则3：If ($c_1$ = 0) ∧ ($c_2$ = 1) then *class* = 1。

共有3条规则，其条件属性个数分别为1、2、2。每条规则的平均条件属性个数为(1 + 2 + 2)/3 =1.67。

**表3 值约简后的决策表**

**Table 3 Value reduction of decision tables**

| *U* | $C_1$ | $C_2$ | $C_3$ | *class* |
|---|---|---|---|---|
| 1 | 1 | ● | ● | 0 |
| 2 | ● | 0 | 0 | 0 |
| 3 | 0 | 1 | ● | 1 |

# 3 实验结果与分析

为验证本文RNDM算法的有效性，将其应用于两个数据库：IBM数据库和Breast Cancer数据库，将结果分别与基于神经网络的数据挖掘方法(NNDM)、基于粗集理论的数据挖掘方法(RSDM)进行了比较。

程序中设$out_i$、$class_i$分别为第$i$个实例的网络输出值和实际决策值，如果$|out_i - class_i|<0.35$，则认为第$i$个实例被正确分类；否则，被认为不能被正确分类。为缩短程序的运行时间，网络的学习算法采用BFGS (Broyden-Fletcher-shanno- Goldfarb) quasiNewton BP算法。交互验证集上分类精确度的最大允许下降Δ$A$=5 %。

## 3.1 IBM数据库

IBM数据库问题中包括9个属性，其属性值的具体意义及产生方法如表4所示。文献[9]使用10个函数在IBM数据库上定义了10个二分类问题。如Function 3定义为：

Class *A*：((*age* < 40) ∧ (*elevel* ∈ [0…1])) ∨

((40 ≤ *age* < 60) ∈ (*elevel* ∈ [1…3])) ∨

((60 ≤ *age*) ∧ (*elevel* ∈ [2…4]))，

Class *B*：otherwise。

**表4 IBM数据库属性值的描述**

**Table 4 Description of IBM database attribute values**

| Attribute | Description | Value |
|---|---|---|
| Salary | Salary | uniformly distributed from 20 000 to 150 000 *salary* ≥ 75 000 → *commission* = 0 else uniformly distributed from 10 000 to 75 000 |
| Commission | Commission | uniformly distributed from 20 to 80 |
| Age | Age | uniformly chosen from 0 to 4 |
| elevel | education level | uniformly chosen from 1 to 20 |
| car | make of the car | |
| zipcode | zip code of the town | uniformly chosen from 9 available zipcodes |
| hvalue | value of the house | uniformly distributed from 0.5*k* 100 000 to 1.5*k*100 000where *k* ∈ {0…9} depends on zipcode |
| hyears | years house owned | uniformly distributed from 1 to 30 |
| loan | total loan amount | uniformly distributed from 1 to 500 000 |

本文按照问题的复杂程度选择Function 3和Function 7两个问题进行实验。每次实验随机产生2 000个实例数据，其中800个实例作为训练集，200个实例作为交互验证集，其余1 000个实例作为测试集。采用温度计编码方法对属性值进行离散化，9个属性转换为37个二进制属性。30次实验的结果如表5所示，其中NNDM列的数据来自文献[5]。

实验中，RNDM算法产生的结果相当稳定。例如，在Function 3问题上，30次实验均得到相同的属性子集{18 ,13 ,11 ,16 ,19}。由表5可以看出，RNDM算法得到的规则其测试精度均高于NNDM算法。这是因为NNDM算法从训练后的神经网络中抽取规则的过程中会对数据产生过拟合，导致其规则的分类精度比神经网络的分类精度要有所下降。而且如前所述，由于粗集理论的使用，RNDM算法的速度要远快于NNDM算法。

**表5 IBM数据库问题的实验结果**

**Table 5 IBM database on the experimental results**

| Function | 算法 | 平均测试精度/% | 平均规则数 | 平均条件属性个数 |
|---|---|---|---|---|
| 3 | RNDM | 99.95(0.01) | 7.07(0.25) | 3.40(0.11) |
| | NNDM | 98.18(1.56) | 6.70(1.15) | 3.18(0.28) |
| 7 | RNDM | 93.72(0.13) | 4.60(1.22) | 1.17(0.37) |
| | NNDM | 90.50(0.92) | 7.43(1.76) | 2.94(0.32) |

注：括号内的数据为标准方差。

## 3.2 Breast 数据库

Wisconsin Breast Cancer Data (Breast)数据库记录了

669 个乳癌病例（实例），每个病例（实例）由 9 项指标（属性）描述，其属性值均为从 1 到 10 的整数。所有实例分属于两类：benign（良性），malignant（恶性）。其中有 16 个病例的指标记录不全，实验中舍弃不用。每次实验中随机选取 366 个实例作为训练集，90 个实例作为交互验证集，其余的 227 个实例为测试集。表 6 为 20 次实验的结果，其中 RSDM 列的数据来自文献[10] (文中没有标明标准方差)。20 次实验中有 13 次得到属性子集 $\{3,6\}$。以 $A_3$，$A_6$ 分别表示第 3 个属性(Uniformity of Cell Size) 和第6 个属性(Single Epithelial Cell Size) 的属性值，Class 表示实例的类，则其中的一组规则为：

Rule 1: if $(A_3 \in \{5...10\})$, then *Class* = *malignant*;

Rule 2: if $(A_6 \in \{6...9\})$, then *Class*= *malignant*;

Rule 3: if $(A_3 \in \{2...5\}$ and $A_6 = 10)$,
then *Class* = *malignant*;

Rule 4: if $(A_3 \in \{3,4\}$ and $A_6 = 3)$,
then *Class* =*malignant*;

Rule 5: if $(A_3 \in \{2,3,4\}$ and $A_6 = 1)$,
then *Class*= *benign*;

Rule 6: if $(A_3 = 1)$, then *Class* = *benign*。

**表 6　Breast 数据库问题的实验**

**Table 6　Experimental database of Breast**

| 算法 | 平均测试精度/% | 平均规则数 | 平均条件属性个数 |
|---|---|---|---|
| RNDM | 94.80(1.34) | 6.14(0.76) | 1.71(0.41) |
| REDM | 92.38 | 7.8 | 1.6 |

注：括号内的数据为标准方差。

## 4　结束语

本文融合粗集理论和神经网络两种技术的优点，提出了一种数据挖掘的新方法。与现有方法相比，该方法具有如下特点：

1) 使用粗集理论和神经网络分两步进行属性约简，粗集理论加快了属性约简的速度，神经网络则提高了属性约简的准确性和稳定性。

2) 使用神经网络过滤数据中的噪声，然后由粗集理论从数据中直接抽取规则，从而在保证规则准确性的同时，避开了从神经网络中抽取规则的困难。实验表明，该方法快速有效，具有良好的鲁棒性，得到的规则知识简单准确，可理解程度高。

**参考文献：**


[1] Chen M S，Han J，Yu P S. Date Mining :An overview from a database perspective[J]. IEEE Trans on Knowledge and Data Engineering，1996，8(6)：866-883.

[2] Bengio Y，Buhmann J M. Introduction to the special issue on neural networks for data mining and knowedge discovery[J]. IEEE Trans on Neural Network，2000，11(3)：545-549.

[3] Pawlak Z，Busse J G. Rough sets[J]. Communciations on the ACM，1995，38(11)：89-95.

[4] Ziarko W. Introduction to the special issue on rough sets and knowledge discovery[J]. Computational Intelligence，1995，11(2)：223-225.

[5] Setiono R，Liu H. Effective data mining using neural networks [J]. IEEE Trans on Knowledge and Data Engineering，1996，8(6)：957-969.

[6] Jelonek J，Krawiec K，Slowinski R. Rough set reduction of attributes and their domains for neural networks[J]. Computational intelligence，1995，11(2)：339-347.

[7] Ahn B S，Cho S S，Kin C Y. The integrated methodology of rough set theory and artificial neural network for business failure prediction[J]. Expert systems with application，2000，18(1)：65-74.

[8] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data[M]. Netherlands：Kluwer Academic Dordrecht，1991.

[9] Bui T，Co-o P. A Group Decision Support System for Cooperative Multiple Criteria Group Decision Making[C]// Lectures Notes in Computer Science. Berlin：Springer，1987：58-59.

[10] Mishra S K，Wang S Y，Lai K K. Optimality and Duality for Multiple-Objective Optimization under Generalized Type I Univexity[J]. Journal of Mathematical Analysis and Applications，2005，303(1)：315-326.