医疗健康数据的模糊粗糙集规则挖掘方法研究

刘 洋 张 卓 周清雷
(郑州大学信息工程学院 郑州 450001)

摘 要 医疗健康数据通常属性较多,且存在连续型、离散型并存的混合数据,这在很大程度上限制了知识发现方法对医疗健康数据的挖掘效率。以模糊粗糙集理论为基础,研究混合数据上的分类规则挖掘方法,通过引入规则获取算法的泛化阈值,来控制获取规则集的大小和复杂程度,提高粗糙集知识发现方法在医疗健康数据上的分类效率。最后通过对比实验验证了该算法在医疗决策表上挖掘规则的有效性。

关键词 电子健康,知识发现,粗糙集理论,规则提取,混合数据
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.12.035

Research on Fuzzy Rough Sets Based Rule Induction Methods for Healthcare Data

LIU Yang ZHANG Zhuo ZHOU Qing-lei
(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract Healthcare databases typically contain numerous attributes, and have both continuous and discrete type of attributes in hybrid data, which limits the mining efficiency of knowledge discovery on healthcare data to a great extent. Based on fuzzy rough sets theory, we studied the classification rule mining methods on hybrid data. By introducing the generalization thresholds for rule induction algorithm, the proposed method can reduce the size of extracted rule set and complexity of rules, which may improve the classification efficiency of rough sets based knowledge discovery method on healthcare data. Finally we conducted comparative experiments on medical decision tables to verify the effectiveness of mined rules of proposed algorithm.

Keywords E-health, Knowledge discovery, Rough sets theory, Rule induction, Hybrid data

1 引言

数据库技术的快速发展使信息技术渗透到包括医疗健康在内的多个领域[1]。医学领域存在大量的数据,比如病人的病史、医生诊断记录、检验报告、临床治疗信息、药品使用信息等[2]。随着区域卫生信息平台中的医疗健康数据库逐渐膨胀、规模逐渐扩大,其复杂度也随之日益增加。尽管积累了大量的医疗健康档案数据,但是很少能将这些数据中有价值的东西挖掘出来用于日常的各种医疗健康决策中。医疗健康档案的建立不仅能从管理上辅助疾病的防治,而且其中还隐含着一些与疾病相关的信息。因此,从医疗健康数据中挖掘出有价值的信息或模式具有重要的意义。

粗糙集理论是由波兰数学家 Pawlak 教授提出的一种处理模糊和不确定知识的数学工具,现已经被广泛应用于数据挖掘、人工智能、模式识别以及分类等诸多领域[3]。粗糙集由于不需要先验知识,因此很适合医学领域的一些应用。为保证知识获取的有效性,医疗健康数据的知识发现方法应遵循简化准则。粗糙集属性约简的目的就是得到其最小的属性子集,但由于该问题已经被证明是 NP-hard 问题,因此现有的研究大都采用启发式搜索策略,获取约简算法的一个或多个可行解。与决策树方法相比,基于粗糙集的规则挖掘方法省去了建树的复杂度,可以挖掘确定规则和可能规则[4]。国内外利用粗糙集理论,挖掘医疗健康数据,提取出隐含在数据内部的对医学实践有意义的信息[5,6]已有一些成功的案例。

医疗健康信息系统中存在多类型属性并存的混合数据,例如由澳大利亚 Garvan 医学研究所收集的甲状腺病症数据包含了两千多个病例的 29 个特征,这些健康数据的属性既包含如性别、是否怀孕、是否有肿瘤等符号属性,也包含了如年龄、甲状腺素结合球蛋白等数值属性。在引入知识发现和数据挖掘方法来处理医疗健康数据时,往往采用离散化算法把数值型变量转化为符号型变量。然而这一转换不可避免地带来了信息损失,学习算法的性能很大程度上取决于离散化的效果。为了解决这一问题,研究者分别引入了模糊粗糙集模型[7]和邻域粗糙集模型[8]。不同的模型基于不同的粒度度量标准和逼近定义,但它们都属于知识发现的研究范畴。胡清华基于模糊信息粒度的信息量度量提出了混合属性约简方法,实验结果表明该方法的约简效率好于基于离散化的属性约简算法[9]。刘金福等基于带权论域上的等价关系,讨论了带权近似空间,并给出有效的规则挖掘方法[10]。但是,还没有文献对面向混合数据的粗糙集规则挖掘方法进行分析和研

到稿日期:2014-03-01 返修日期:2014-04-08 本文受国家自然科学基金项目(61303044),郑州市科技攻关项目(131PPTGG409-30)资助。
刘 洋(1984—),男,博士,讲师,主要研究方向为粗糙集、模糊集和电子健康,E-mail:ieyangliu@zzu.edu.cn;张 卓(1978—),男,博士,讲师,主要研究方向为数据挖掘、模糊概念格;周清雷(1962—),男,教授,博士生导师,主要研究方向为自动机理论、信息安全、智能系统。

究。本文将系统地研究基于模糊粗糙集模型的规则挖掘算法，该方法省去了混合型属性的预处理过程，可以直接分析混合数据，为使用粗糙集方法在医疗健康数据上发现分类知识提供了支撑。

2 基于粗糙集的 LERS 规则挖掘系统

基于粗糙集的 LERS 规则挖掘系统是一种目前应用最为广泛的粗糙集知识获取系统。LEM2 算法是 LERS 系统的典型规则挖掘算法，Pawlak 粗糙集模型的上、下近似集作为算法的输入，为不一致数据集生成确定规则和可能规则。LEM2 算法已被成功应用在医疗决策表的知识发现和数据挖掘中[4]。

2.1 粗糙集规则挖掘模型

LEM2 算法的基本概念是利用属性值对进行建模。对于一个属性值对 $t=(c_j,v),c_j\in C,v\in V,t$ 的块记为$[t]$，它是由论域 U 中的一系列对象组成，这些对象满足$\{x\mid x_j=v,x\in U\}$。设 B 是一个类别集合的非空下、上近似集合，T 是一个对象的属性值对集合，称集合 B 依赖于集合 T 当且仅当 $\varnothing\neq[T]=\bigcap\limits_{t\in T}[t]\subseteq B$。集合 T 称为集合 B 的最小覆盖，当且仅当 B 依赖于 T，并且不存在 T 的其它子集 T'，满足 B 依赖于 T'。设$\mathbb{C}$是一个非空的最小覆盖集，$\mathbb{C}$为 B 的一个局部覆盖当且仅当：

1) $\bigcup\limits_{T\in\mathbb{C}}[T]=B$；

2)$\mathbb{C}$为最小，即不存在$\mathbb{C}$的子集$\mathbb{C}'$满足条件 1)。

LEM2 算法采用基于优先级的策略每次选择具有最高优先级别的属性值对，并将其加入到部分最小覆盖中。当得到一个局部覆盖时，算法将其转化为一组规则集。基于粗糙集的规则挖掘方法不需要进行模型上的改动，可以与离散化方法、属性约简方法相结合进行医疗健康混合数据上的知识获取。

2.2 分类规则的衡量标准

Pawlak 给出了广义决策规则的性质。给定一个决策信息系统 DIS,D_K 为一个广义决策类，其中 K 表示与 D_K 相关的一个用例子集。p 表示一个基本条件，即形如 (a,v)的条件属性-值对，其中 $a\in C$ 且 $v\in V_a$，设 $\Phi=p_1\wedge p_2\wedge\cdots\wedge p_n$ 为一个基本条件的合取式，$[\Phi]$表示 Φ 的覆盖，即满足 Φ 的所有基础条件的用例子集，$[\Phi]_K^{+}=[\Phi]\cap K$ 称为 Φ 在 K 上的正覆盖，$[\Phi]_K^{-}=[\Phi]\cap(U\backslash K)$称为 Φ 在 K 上的负覆盖。

定义 1 给定一个决策表 DIS,D_K 为一个广义决策类，分类规则 r 可表示为：

IF Φ THEN D_K，

或简记为：$\Phi\rightarrow D_K$

其中，Φ 称为 r 的条件部分，满足$[\Phi]_K^{+}\neq\varnothing$，且 D_K 称为 r 的决策部分。若 D_K 为单个决策，规则 r 称为确信规则；若 D_K 为多个决策的联合，规则 r 称为可能规则。

定义 2 给定一个分类规则 r,K 为一个广义决策 D_K 覆盖的用例子集，分类规则 r 相对于 K 是显著的当且仅当满足以下条件：

(1)一致性：$[\Phi]_K^{-}\neq\varnothing$；

(2)最小的：若从 Φ 中删除任意一个基本条件 p_i，则将不再满足一致性。

定义 3 给定一个分类规则集合 R,D_K 为一个广义决策类，如果 R 对 D_K 的描述满足以下条件，则称 R 是一个最小规则集。

(1)对于任意的 $r\in R$ 是显著的；

(2) $\bigcup\limits_{r\in R}[\Phi]=K$；

(3)删除任意一个规则 $r\in R$ 后，$R\backslash\{r\}$不再满足条件(1)和(2)。

定义 4 给定一个决策信息系统 DIS，r 为决策表的一个分类规则，Φ 为规则的条件部分，D_K 为规则的决策部分，则规则的支持度、覆盖度和相对某一决策 $d_i\in V_d$ 的可信度分别定义为：

$\mu_{sup}(r)=|[\Phi]|/|U|$

$\mu_{cov}(r)=|[\Phi]|/|D_K|$

$\mu_{cer}(r,d_i)=|[\Phi]_{d_i}^{+}|/|[\Phi]|$

由定义 4 可知，支持度表示论域中支持规则的用例数；覆盖度表示规则的支持数在相应的广义决策类中的比重；可信度表示运用该规则进行推理正确的概率。当分类规则获取后就需要对被提取的规则进行评价。通常的方法是检验规则对测试样本分类的识别率，错误率越小的规则越好。

3 面向医疗健康混合数据的规则挖掘算法

Pawlak 粗糙集模型仅工作在含有符号型属性的数据集上。医疗决策表通常含有异质类型的属性。下面将以模糊粗糙集模型为基础讨论混合数据上的分类规则挖掘算法。

3.1 基于模糊粗糙集的规则挖掘模型

一个清晰的等价关系可以生成论域上清晰的划分，而一个模糊等价关系可以生成论域上的模糊划分。因此，粗糙集规则挖掘算法中属性值对的块定义可以很自然地推广到模糊等价空间。

定义 5 设 $t=(c_j,F)$是混合决策信息系统上的属性值对，则 t 的模糊块定义为：

$$[t]=\int_{x\in U}\mu_F(x_j)$$

显然，$[t]$是模糊等价关系 R 上由 x 生成的模糊等价类。根据模糊等价关系的性质，$[t]$是一个模糊集合。这是模糊块和清晰块的主要区别。很容易看出，正如模糊集合是清晰集合的很自然的扩展，模糊块的定义也是对清晰块定义的很自然的扩展。如果属性是离散型的，等价关系矩阵和等价类将退化为经典算法的块定义。

定义 6 设 R 为模糊等价关系，X 为对象的清晰集合，则 X 的下、上近似集定义为：

$\underline{R}X=\{x\mid[x]_R\subseteq X,x\in U\}$

$\overline{R}X=\{x\mid[x]_R\cap X\neq\varnothing,x\in U\}$

模糊集合理论中的包含操作“$\subseteq$”首先由 Zadeh 教授提出，称为 Zadeh 包含。然而，该定义在实际应用中过于严格。这里采用一种弱包含操作符“$\subseteq_\alpha$”，即 $\forall x\in U,A\subseteq_\alpha B$ 当且仅当 $\max(1-\mu_A(x),\mu_B(x))\geqslant\alpha$。

定义 7 设 X 是一个清晰集合，T 是一个属性值对的集合，集合 X 以 α-近似依赖于集合 T 当且仅当：

$$\varnothing\neq[T]=\bigcap_{t\in T}[t]\subseteq_\alpha X$$

其中，$A\cap B=\int_{x\in U}\min(\mu_A(x),\mu_B(x))/x$。

设 B 是一个类别的非空下、上近似的清晰概念集，T 是一个属性值对集合。集合 T 是 B 的一个近似最小覆盖当且仅当 B 以 α-近似依赖于 T，并且不存在 T 的子集 T'使得 B 满

足 α-近似依赖于 T'。

定义 8 设$\mathbb{C}$是一个非空的近似最小覆盖集，即属性值对集合的集合，B 是论域上的一个类别的清晰对象集合。$\mathbb{C}$ 是 B 的一个(α,k)-近似局部覆盖当且仅当满足下列条件：

1) 对于$\mathbb{C}$的任意元素 T，T 是一个 B 的一个 α-近似最小覆盖；

2) $H(\bigcup_{T\in\mathbb{C}}[T],B)\leqslant k$；

3)$\mathbb{C}$是最小的，即$\mathbb{C}$具有最少的元素数量。

其中，两个模糊集合的相异度度量定义为：

$$H(A,B)=\frac{|\overline{A}\cap B|+|A\cap\overline{B}|}{|A\cup B|},|\cdot|=\sum_{x\in U}\mu_{\cdot}(x),$$

$$\overline{A}=\int_{x\in U}(1-\mu_A(x))/x,H(A,B)\in[0,1]$$

由定义可知，α 和 k 为规则挖掘算法提供了新的停止准则。当 $\alpha=0,k=0$ 时，属性为离散型变量，停止准则退化为与算法 LEM2 相同的停止条件。由于算法 LEM2 的停止条件过于严格，它可能提取过于细化的分类规则，导致获取的知识过拟合训练数据，并且算法的运行时间也因此急剧增加。为解决这一问题，将 α 和 k 引入到规则挖掘算法中，允许输入集合部分依赖于概念集，使局部覆盖可以不包含一小部分的训练数据。参数 α 和 k 起到了松弛规则挖掘算法的作用，当数据集含有分布重叠的类或者不一致对象时，这种机制非常有效。

3.2 混合数据上的模糊粗糙集规则挖掘算法 FRLEM2

算法 LEM2 利用启发式信息提取频繁的属性值对来组成最小覆盖，与 LEM2 类似，算法 FRLEM2 根据评分函数的定义每次选择具有最高评分的属性值对来形成近似最小覆盖，通过参数 α 和 k 有效控制生成的规则集的复杂程度，提高获取知识的泛化能力。由于模糊块的定义与经典块不同，需要对属性值对的评分函数重新定义。

定义 9 给定属性值对 t 和模糊集合 G，t 相对于 G 评分函数定义为：

$$Score(t,G)=|[t]\cap G|$$

设 $HDIS=\{U,C^s\cup C^n\cup C^f\cup\{d\},V,f\}$为混合决策表，其中 C^s 表示符号型属性集合，C^n 表示数值型属性集合，C^f 表示模糊型属性集合，d 表示决策属性。B 表示训练数据集中某一类别的清晰下、上近似集。α 和 k 为两个通常设置在零附近的算法参数。下面给出了分类规则挖掘算法 FRLEM2 的代码。

算法 基于模糊粗糙集模型的 FRLEM2 规则挖掘算法

输入：混合决策表 HDIS，一个非空概念集 B，参数 α 和 k

输出：一个(α,k)-近似局部覆盖$\mathbb{C}$

```
1. G⇐B;
2. ℂ⇐∅;
3. while H(⋃_{T∈ℂ}[T],B)>k do
4.     T⇐∅;
5.     while (T=∅) or (not([T]⊆_αB)) do
6.         t⇐arg max_{t∈T} Score(t,G);
7.         T⇐T∪{t};
8.         G⇐[t]∪G;
9.     end while
10.    for ∀t∈T do
11.        if T\{t}⊆_αB and T\{t}≠∅ then
12.            T⇐T\{t};
13.        end if
14.    end for
15.    ℂ⇐ℂ∪{T};
16.    G⇐B\⋃_{T∈ℂ}[T];
17: end while
18: for ∀T∈ℂ do
19.    if H(⋃_{S∈ℂ\{T}}[S],B)⩽k then
20.        ℂ⇐ℂ\{T};
21.    end if
22. end for
23. return ℂ.
```

定理 1 当 $k=\alpha=0$，且属性全为连续型属性时，分类规则挖掘算法 FRLEM2 退化为算法 LEM2。

证明：如果 $k=\alpha=0$，且属性全为连续型的，算法 FRLEM2 的以下性质成立：

1)模糊等价关系退化为清晰的等价关系；

2)α-依赖关系定义等价于经典的依赖关系；

3)(α,k)-局部覆盖等价于经典的局部覆盖；

4) 模糊集合操作退化为经典集合的操作。

由 1)、2)、3)和 4)可知，算法 FRLEM2 和 LEM2 的规则挖掘过程的停止准则等价，因此定理成立。

定理 1 说明了算法 LEM2 是算法 FRLEM2 在 $k=0,\alpha=0$，且属性全为离散型时的特例。事实上，算法 FRLEM2 避免了过于严格的终止准则以及 LEM2 算法对属性类型的限制。算法 FRLEM2 的性能也可以由参数 α 和 k 进行调节。

4 医疗健康数据上的实验与结果分析

4.1 2-型糖尿病数据的知识发现实例

医学诊断知识挖掘系统的实现有助于 2-型糖尿病并发症发病规律的研究。本节将算法 FRLEM2 应用于一个 2-型糖尿病并发症数据集的诊断规则自动发现。算法 FRLEM2 利用模糊 C 平均算法将连续型属性模糊化到 3 个区间，模糊集合的隶属函数采用等腰梯形分布，区间端点作为等腰梯形分布中参数 a,b 的取值。为统一起见，设置算法 FRLEM2 的参数 α 和 k 分别为 0.02 和 0.025。

实验用到的数据集来源于医疗机构对河南安阳地区 16302 个人的随机抽样健康体检调查。每个样本分别描述个人资料、生活和工作状况、既往家族史、体检结果和实验室检验数据 5 个部分。个人资料部分主要是被调查人群的一般情况，包括年龄、性别、民族、婚姻、学历；生活和工作状况部分主要是为了研究糖尿病与个人生活和工作状况之间的关系，主要包括职业、抽烟、喝酒、锻炼、睡眠时间、饮食等方面；家族病史调查部分主要是为了研究糖尿病与相关家族病史之间的关系，调查的相关家族病史主要包括高血压、糖尿病、高血脂、心脏病和脑血管病，范围包括父母、兄弟姐妹、子女、姑叔、姨舅等亲属；体检数据包括身高、体重、腰围、臀围及血压值；实验室检查数据包括胆固醇、甘油三脂、低密度脂蛋白、高密度脂蛋白和血糖等。数据集中每个样本含有 52 个条件属性和一个决策属性，条件属性中含有符号型属性 14 个，数值型属性 38 个。样本的年龄从 45 岁到 96 岁，其中仅有 538 个样本的年龄小于 53 岁。所有样本的平均年龄是 62 岁。其中 2-型糖尿病慢性并发症的患病人群占 28%。数据集中 10000 个样本用作训练集，另外 6302 个样本用作测试集。

利用混合数据属性约简算法对训练集进行约简，得到的前 6 个医学诊断征兆依次为血糖值 GLU、糖尿病家族史 DH、

饮食习惯 DIET、甘油三酯 TG、年龄 AGE 和工作状况 JOB。针对获取的约简属性集，分别应用粗糙集规则提取算法 LEM2 和算法 FRLEM2 进行诊断规则挖掘。两种挖掘算法在约简属性集上的分类精度和获取的规则数见表 1。

表 1　LEM2 和 FRLEM2 算法在医疗数据上的实验结果

算法	分类精度	分类规则数
LEM2	0.87	33
FRLEM2	0.91	8

结果表明，两种规则挖掘算法的分类精度都较理想，但算法 FRLEM2 的分类精度和规则简洁程度都好于 LEM2。算法 FRLEM2 在前 6 个属性集上获取的确信规则集合见表 2。

表 2　FRLEM2 算法挖掘的分类规则集

规则	支持度	可信度
IF GLU is TP(7.0,17.44) THEN D=Y	0.27	1
IF GLU is TP(5.85,7.0) and JOBSTR=L and DH=Y THEN D=Y	0.04	0.97
IF GLU is TP(5.85,7.0) and AGE is TP(55,86) and TG is TP(1.73,7.43) THEN D=Y	0.17	0.58
IF GLU is TP(5.85,7.0) and DIET=H and DH=Y THEN D=Y	0.18	0.95
IF GLU is TP(0,5.85) and AGE is TP(45,55) THEN D=N	0.11	1
IF GLU is TP(0,5.85) and DIET=L THEN D=N	0.02	1
IF GLU is TP(5.85,7.0) and DIET=M and TG is TP(0,1.73) THEN D=N	0.06	0.96
IF GLU is TP(5.85,7.0) and JOBSTR=H THEN D=N	0.14	1

从规则中可以明显看出，由算法 FRLEM2 提取的规则集具有较高的支持度和可信度，揭示了血糖值、年龄、家族史等因素在 2-型糖尿病发病规律中的重要性，这与医学上有关 2-型糖尿病发病规律的研究基本一致。

4.2　规则挖掘算法的对比实验分析

下面通过实验将本文算法与经典知识获取方法 C4.5、前端配有离散化 FCM 模块的 LEM2 算法[4]在医疗决策表上的规则挖掘效率进行对比测试。为测试各算法的有效性，从美国加州大学 Irvine 分校建立的机器学习测试数据库下载了数据集：Heart，Cleve 和 Hypothyroid。各数据集的描述见表 3。实验对数据集进行 10 次交叉验证。对比实验选择的医疗健康数据集都包含了混合型属性。

表 3　对比实验的数据集描述

缩写	数据集	用例数	属性数	离散属性	连续属性
hea	StatLog heart disease	270	13	7	6
cle	Cleve database	303	14	6	8
hyp	Hypothyroid disease	3163	25	7	18

各算法在医疗决策表上的分类精度结果见表 4。表中记录了算法获取的分类精度平均值，由于采用十次交叉验证实验，因此结果包含了各实验的平均值和标准方差。由分类精度对比结果可以看出，算法 C4.5 和 FRLEM2 在所有数据集上都具有较好的平均分类精度。仔细观察表 4 可以发现，没有任何分类器在所有数据集上能够得到最优分类精度结果。

表 4　分类精度对比结果

数据集	C4.5	LEM2-FCM	FRLEM2
hea	77±8.0	79±6.7	83±7.6
cle	55±10.3	56±13.2	58±13.7
hyp	99±3.9	83±8.6	94±3.8
平均	77	73	78

各算法生成的规则集结果见表 5。该表给出了算法生成规则的数量 r 和规则的平均长度 l/r，规则的数量可以视为规则集的复杂程度，而规则的平均长度可以更加清楚地反映平均每条规则的简洁程度。

表 5　分类规则集对比结果

数据集	C4.5		LEM2-FCM		FRLEM2	
	r	l/r	r	l/r	r	l/r
hea	27	3.7	36	2.3	17	2.4
cle	21	2.8	38	3.5	25	2.6
hyp	7	3.9	46	4.1	14	4.5
平均	18	3.5	40	3.3	19	3.2

由表 5 可以看出，算法 FRLEM2 生成了平均数量最少的规则集。由此可以得出结论，算法 FRLEM2 是对 LEM2 算法的重大改进。由于各算法规则的平均长度都较小且较接近，因此各算法在基准测试实验中都可以生成单个规则较为简洁的结果。

结束语　基于粗糙集的医疗健康数据知识发现方法在不需要任何先验知识的前提下，就能有效地挖掘医疗健康决策表。然而，粗糙集规则挖掘方法只适用于离散型数据，对于实际医疗领域中大量的离散型、连续型属性并存的医疗健康混合数据不能进行直接有效的挖掘。本文研究了医疗健康混合数据上的模糊粗糙集分类规则挖掘方法，该方法兼顾规则挖掘结果的性能和信息融合应具备的功能，通过设置合适的参数 α 和 k 来寻找近似局部覆盖。实验结果表明，模糊粗糙集分类规则挖掘方法提高了粗糙集规则挖掘算法在医疗健康混合数据上的分类精度，同时控制了生成规则的数量。

参 考 文 献

[1] 吴信东，叶明全，胡东辉，等. 普适医疗信息管理与服务的关键技术与挑战[J]. 计算机学报，2012，35(5)：827-845

[2] Koh H C, Tan G. Data mining applications in healthcare[J]. Journal of Healthcare Information Management, 2011, 19(2): 65

[3] 王国胤，姚一豫，于洪. 粗糙集理论与应用研究综述[J]. 计算机学报，2009，32(7)：1229-1246

[4] Grzymala-Busse J W, Hippe Z S, Piatek L. A System for Melanoma Diagnosis Based on Data Mining[J]. Medical Applications of Artificial Intelligence, 2013: 165

[5] 鄂旭，邵良杉，张毅智，等. 一种基于粗糙集理论的规则提取方法[J]. 计算机科学，2011，38(1)：232-235

[6] 杨宏薇，何中市. 粗糙集属性约简方法及其在医疗中的应用研究[J]. 计算机工程与应用，2010，46(25)：207-210

[7] Hu Q, An S, Yu X, et al. Robust fuzzy rough classifiers [J]. Fuzzy sets and systems, 2011, 183(1): 26-43

[8] Zhu P, Hu Q. Adaptive Neighborhood Granularity Selection and Combination Based on Margin Distribution Optimization[J]. Information Sciences, 2013, 249(10): 1-12

[9] Hu Q, Xie Z, Yu D. Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation[J]. Pattern Recognition, 2007, 40(12): 3509-3521

[10] Liu J, Hu Q, Yu D. A weighted rough set based method developed for class imbalance learning [J]. Information Sciences, 2008, 178(4): 1235-1256