文章编号:1001 - 9081(2008)06 - 1420 - 04

广义粗糙集理论及实值属性约简

肖 迪[1], 张军峰[2]
(1. 南京工业大学 自动化学院,南京 210009; 2. 南京航空航天大学 自动化学院,南京 210016)
(xiaodi_12@sina.com)

摘 要:针对经典粗糙集理论仅能处理离散化数据的局限性,提出属性和属性子集的广义重要度的概念以及空间中的广义近邻关系,并提出了广义近邻关系下的广义粗糙集扩展模型。广义粗糙集理论利用广义近邻关系在全局中划分相容模块,构成集合的下、上近似集,避免了经典粗糙集理论必须量化数据的麻烦。另外,提出了广义粗糙集的实值属性约简的一种贪心算法,并分析了约简属性集合的质量。最后通过实例验证了所提方法的正确性和有效性。

关键词:数据挖掘;广义粗糙集理论;广义重要度;近似约简
中图分类号: TP18 **文献标志码**:A

Generalization rough set theory and real-valued attributes reduction

XIAO Di[1], ZHANG Jun-feng[2]
(1. *College of Automation, Nanjing University of Technology, Nanjing Jiangsu* 210009, *China*;
2. *College of Automation, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu* 210016, *China*)

Abstract: Considering that the classical rough set theory can only process the discrete data, the degree of general importance of an attribute and attribute subsets was presented. And then a generalization rough set theory was proposed based on the general near neighborhood relation. The theory partitioned the universe into the tolerant modules and formed lower approximation and upper approximation of the set under general near neighborhood relationship which avoided the discretization in Pawlak's rough set. Furthermore, the definition of attribute reduction in generalization rough set and its greedy algorithm were proposed. Finally, results of some examples show the correctness and validity of this method.

Key words: data mining; general rough set theory; degree of general importance; approximation reduction

0 引言

一般来说,对一个确定的故障系统或一个数据库,它们所包含的信息总会有冗余和不必要的情形存在。如果把每个信息看作一个特征向量,那么,从原始特征集合中找出一系列候选特征,就是特征选择。粗糙集理论的属性约简是一种特征选择方法[1]。粗糙集属性约简可以从众多的特征集合中找出最简约或近似最简的特征子集,并且这些特征子集不会影响全局决策分类的结果。粗糙集理论的属性约简是以最小概念描述为准则来选择特征[2],即它可以选择最小的特征子集,但仍然能完全地描述全局所有对象的特性。

粗糙集属性约简的主要障碍是必须计算可区分函数和正域,这些计算均以量化决策表为基础开展[3]。因此对于连续属性值决策系统必须首先进行离散化处理,而离散化过程中断点集的选取是离散化效果优劣的关键[4]。在实际系统中,噪声和干扰是无法避免的。断点所选择的位置若引起噪声数据的离散化结果的差异,势必也影响问题的解决[5]。为了减小噪声数据在离散化过程中所产生的影响,或者说增加粗糙集理论对噪声数据的容错能力,文献[6]提出了可变精度粗糙集理论,通过一个模型参数调节可允许的不确定水平来定义每个集合的粗糙集概念。因此,可变精度粗糙集模型增强了粗糙集理论的实际应用能力。文献[7]提出相容关系代替不可分辨关系,可以提高粗糙集方法的适用性。文献[8]建立了相容粗糙集模型,在不需要完全等价分类的信息修复和文本检索方面取得了很好的效果。

本文提出一种扩展的粗糙集理论,对于属性值集为连续实数的决策系统,不需要经过离散化过程,而根据属性的特征定义了属性的广义重要度,从而可以重新度量空间中样本之间的相似性距离,然后以广义欧氏距离为基础构成了空间中的广义近邻关系,在此关系下定义了广义近邻关系粗糙集并研究了实值属性的约简方法。

1 属性的广义重要度

1.1 基本概念

设决策系统 $S = \langle U, A\rangle$, $A = C \cup \{d\}$, U 为全体元素, C 为条件属性集, d 为决策属性集。对任意属性 $a \in A$, 有 $a: U \to V_a$, 其中 V_a 为属性 a 的值域集。设 $d = \{1, 2, \cdots, r(d)\}$, $r(d)$ 为决策类的数目;决策 d 对全局的划分为 $U/d = \{D_1, D_2, \cdots, D_{r(d)}\}$, D_i 为第 i 个决策类, $D_i = \{x \in U \mid d(x) = i\}$。

设 $X \subseteq U$, 粗糙集理论中根据属性子集 $B(B \subseteq C)$ 确定的下、上近似集表示为 $(\underline{B}(X), \overline{B}(X))$。$POS_B(X) = \underline{B}(X)$ 称为 X 的 B 正域,它表示根据属性子集 B, U 中所有一定能归入集合 X 的元素构成的集合。

1.2 经典粗糙集的属性重要度

在经典粗糙集理论中,属性可约简的定义是若去除一个属性,剩余属性对于确定性决策分类的影响不变,则该属性可约简。而对应的属性的重要度定义为去除该属性或属性子集后对确定性决策分类的影响程度,表示为:

收稿日期:2007 - 12 - 21;**修回日期**:2008 - 02 - 02。

作者简介:肖迪(1975 -),女,河北大城人,讲师,博士,主要研究方向:机器学习、图像处理、模式识别; 张军峰(1979 -),男,江苏南通人,博士研究生,主要研究方向:人工智能、故障预报。

$$\sigma(\hat{A}) = 1 - \frac{\sum_{i=1}^{r(d)} |POS_{C-\hat{A}}(D_i)|}{\sum_{i=1}^{r(d)} |POS_C(D_i)|} \tag{1}$$

其中 $\hat{A}$ 为任意单个属性 a 或属性子集 B,$|\cdot|$ 表示集合中元素的个数。

经典粗糙集理论中的属性重要度与其属性约简所针对的对象是一致的,即都是针对确定性决策的元素来考虑的。并且有这样的事实成立:在经典粗糙集理论中,属性重要度的值为零是该属性可约简的必要而非充分条件。

1.3　属性和属性子集的广义重要度

经典粗糙集理论的属性重要度研究对确定性决策元素的影响程度,下面提出一种更为广义的属性重要度,它针对全局中所有元素的影响程度。

命题1　设 $\forall a \in C$,当且仅当 $a(D_i) \cap a(D_j) = \varnothing(\forall i, j = 1,2,\cdots,r(d), i \neq j)$ 时,$U/a = U/d$。

其中 $a(D_k) = [\min(a(x)), \max_{x \in D_k}(a(x))]$,$a(D_i) \cap a(D_j) \subseteq W_a$ 表示属性 a 对应决策类 D_i 的属性值子集与对应决策类 D_j 的属性值子集的交集部分。

命题1 说明,如果一个属性对应不同的决策的值域集是互不相交的,则这个属性与决策属性的划分完全一致。因此,这个属性就是最重要的属性之一,是属性集的核。若某属性在各类区间上的值集是重叠的,即各类决策对应的属性值集是

$$\sigma_g(B) = \begin{cases} 1 - \dfrac{1}{C_{r(d)}^2} \displaystyle\sum_{i \neq j, i,j=1}^{r(d)} \dfrac{B(D_i) \cap B(D_j)}{\max\limits_{\forall k1,k2=1,\cdots,r(d)} (B(D_{k1}) \cap B(D_{k2}))} = 1 - \dfrac{1}{C_{r(d)}^2} \displaystyle\sum_{i \neq j, i,j=1}^{r(d)} \prod_{t=l(1)}^{l(m)} \dfrac{a_t(D_i) \cap a_t(D_j)}{\max\limits_{\forall k1,k2=1,\cdots,r(d)} (a_t(D_{k1}) \cap a_t(D_{k2}))}, & \exists i, j, B(D_i) \cap B(D_j) \neq \varnothing \\ 1, & \text{其他} \end{cases} \tag{3}$$

其中 $C_{r(d)}^2$ 意义同上,$B = \{a_{l(1)}, a_{l(2)}, \cdots, a_{l(m)}\}$,$B(D_i) \cap B(D_j)$ 表示属性子集 B 对应的决策类 D_i 所包围的属性值区域与它对应的决策类 D_j 所包围的属性值区域的交集的面积,$\max_{\forall k1,k2=1,\cdots,r(d)} (B(D_{k1}) \cap B(D_{k2}))$ 表示属性子集 B 对应全部两两决策所有相交部分所包围的最大区域的面积。

定义1中的属性的广义重要度与式(1)的属性重要度是不同的。从上面可知,在经典粗糙集中,如果属性是可约简的,则该属性重要度的值为零,但它的属性广义重要度不一定为零。

例:设实数值决策表如表1所示,比较各属性和属性子集的重要度和广义重要度。

表1　实数值决策表

U	a_1	a_2	a_3	d
x_1	1.6	0.4	5.7	1
x_2	2.2	0.6	6.8	1
x_3	1.9	0.4	5.3	2
x_4	3.4	0.6	4.8	2
x_5	2.0	0.4	6.9	3
x_6	2.9	0.6	7.5	3

对表1所示的决策表,$U/d = \{D_1, D_2, D_3\} = \{\{x_1, x_2\}, \{x_3, x_4\}, \{x_5, x_6\}\}$。

根据定义1、2分别计算表1中所有单个属性和属性子集的广义重要度。采用 Rosetta 工具箱中的 Navie algorithm 对表1进行离散化处理,之后按照式(1)计算各属性和属性子集的重要度,计算结果如表2所示。

从表2中可以看出,在本例中,采用 Navie algorithm 离散化的方法得到的单个属性重要度值都为零,因此无法区别它们对于分类的影响,只能在 a_1, a_3 的属性子集重要度的计算时,可以看出属性 a_1, a_3 是重要的。而属性的广义重要度不仅很好地度量了每个属性分类能力,并且对影响分类的程度的大小更为明确。

表2　各属性及其属性子集的重要度与广义重要度值比较

属性重要度	a_1	a_2	a_3	a_1a_2	a_1a_3	a_2a_3	$a_1a_2a_3$
σ	0	0	0	0	1	0	1
σ_g	0.533	0	1	0.533	1	1	1

定义1　给定决策系统 $S = \langle U, C \cup d \rangle$,$\forall a \in C$,定义属性 a 的广义重要度为:

$$\sigma_g(a) = \begin{cases} 1 - \dfrac{1}{C_{r(d)}^2} \displaystyle\sum_{i \neq j, i,j=1}^{r(d)} \dfrac{a(D_i) \cap a(D_j)}{\max\limits_{\forall k1,k2=1,\cdots,r(d)} (a(D_{k1}) \cap a(D_{k2}))}, & \exists i, j, a(D_i) \cap a(D_j) \neq \varnothing \\ 1, & \text{其他} \end{cases} \tag{2}$$

其中 $C_{r(d)}^2$ 表示从 $r(d)$ 个数中取 2 的组合,$\max_{\forall k1,k2=1,\cdots,r(d)} (a(D_{k1}) \cap a(D_{k2})) \subseteq W_a$ 表示属性 a 对应全部两两决策类的属性值子集的所有交集所包围的最大区间,即 $\max_{\forall k1,k2=1,\cdots,r(d)} (a(D_{k1}) \cap a(D_{k2})) = [\min(a(D_{k1}) \cap a(D_{k2})), \max_{\forall k1,k2=1,\cdots,r(d)} (a(D_{k1}) \cap a(D_{k2}))]$。

由定义1可知,$\sigma_g(a) \in [0,1]$。当所有组合 $a(D_i) \cap a(D_j) = \varnothing$ 时,$\sigma_g(a) = 1$,此时属性 a 的广义重要度最大;当所有组合 $a(D_i) = a(D_j)$ 时,$\sigma_g(a) = 0$,此时属性 a 的广义重要度最小。

定义2　给定决策系统 $S = \langle U, C \cup d \rangle$,$\forall B \subseteq C$,则属性子集 B 的广义重要度为:

定理1　给定决策系统 $S = \langle U, C \cup d, V \rangle$,若 $B_1 \subseteq B_2 \subseteq C$,则 $\sigma_g(B_1) \leqslant \sigma_g(B_2)$。

证明　对 $\forall a_t \in C$,有 $\dfrac{a_t(D_i) \cap a_t(D_j)}{\max\limits_{\forall k1,k2=1,\cdots,r(d)} (a_t(D_{k1}) \cap a_t(D_{k2}))} \leqslant 1$,且根据属性子集重要度的定义,如果 $B_1 \subseteq B_2$,则 $\sigma_g(B_1) \leqslant \sigma_g(B_2)$。

命题2　对一决策系统,如果它的属性集的广义重要度的值为1,则该系统中的决策类是线性可分的,否则就是线性不可分的。

实际上,决策系统属性集的广义重要度若为1,表明应用该属性集可以为全局进行确定性的分类。如果属性集的广义重要度不为1,则该决策系统中的各决策类之间存在交集,这些相交的部分就不能确定分类,因此可以利用粗糙集理论进行研究。

2　广义粗糙集理论及其性质

设 U 为非空、有限的集合,称做论域;C 为条件属性集合 $C = \{a_1, a_2, \cdots, a_m\}$,并且属性值集都是连续的实数值。

定义3 设任意 $x_i, x_j \in U$, B 为 C 的一非空有限属性子集，则 x_i, x_j 在属性集 B 下的广义重要度欧氏距离为：

$$d_g^B(x_i, x_j) = \sqrt{\sum_{a_i \in B} \sigma_g(a_i)(x_i^{(a_i)} - x_j^{(a_i)})^2}$$

简称为广义欧氏距离。一般可记做 $d_g(\cdot, \cdot)$。如果下面的不等式成立：

$$d_g^B(x_i, x_j) < \delta$$

其中 δ 为正实数，则称 x_j 是 x_i 在 B 下的一个 δ 广义近邻，也可以说 (x_i, x_j) 满足 δ 广义近邻关系，记做 $\delta_B(x_i, x_j)$。

性质1 设 $x, y, z \in U$，且设 $d(x,y) > d(x,z)$，但是 $d_g(x,y) > d_g(x,z)$ 不一定成立。

$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$，表示元素 x, y 之间的欧氏距离。

性质1说明广义重要度欧氏距离为了更好地分类，可以改变空间中的距离关系。虽然广义欧氏距离有可能与欧氏距离存在不同的拓扑特性，但是在实际应用中，这种情况不会存在。因为为了达到一致的分类结果，属性约简的结果和属性的广义重要度不会使最后计算出的广义欧氏距离与欧氏距离偏差太远。而广义欧氏距离根据各属性的重要程度计算元素之间的距离，同欧氏距离相比更具有适应性。

把 x 在 U 中的所有属性子集 B 下的 δ 广义近邻称做 x 的 B-广义近邻类，记做 $[x]_B^\delta$，即 $[x]_B^\delta = \{y \in U \mid \delta_B(x,y)\}$。

定义4 设 U 为论域，$X \subseteq U$，A 为全体属性集合，$B \subseteq A$，则 X 在属性子集 B 下的 δ- 下近似集和 δ- 上近似集分别为：

$$\underline{B}(X)^\delta = \{x \in U \mid [x]_B^\delta \subseteq X\}$$

$$\overline{B}(X)^\delta = \{x \in U \mid [x]_B^\delta \cap X \neq \varnothing\}$$

X 在属性子集 B 下的 δ- 上、下近似集之差称做 X 在属性子集 B 下的 δ- 边界集，记做 $BN_B^\delta(X)$，即 $BN_B^\delta(X) = \overline{B}(X)^\delta - \underline{B}(X)^\delta$。

δ 广义近邻关系下的粗糙集理论我们简称为广义粗糙集理论。

3 广义粗糙集的属性约简

定义5 给定决策系统为 $S = \langle U, C \cup d \rangle$，设 $B \subseteq C$，若满足下列条件：

1) $\sigma_g(C) - \sigma_g(B) \leqslant \varepsilon$；

2) $\forall B' \subset B, \sigma_g(C) - \sigma_g(B') > N\varepsilon$；

则称 B 为 C 的广义粗糙集近似约简[9]，记做 $RED_g^\delta(C)$。其中 ε 为一任意小的正数，$N \geqslant 2$，为正整数。

命题3 广义粗糙集理论中，最小 δ 约简的计算是 NP-hard 问题。

证明 已知经典粗糙集理论中的找到最小属性约简的计算是 NP-hard 问题，即它等同于一个 polynomial-time 问题。广义粗糙集理论的最小 δ 约简的计算类似于粗糙集的最小约简问题，因此它也是一个 NP-hard 问题。

命题4 广义粗糙集理论中，广义重要度为零是该属性可 δ 约简的充分而非必要条件。

由命题4可看出，广义粗糙集的广义重要度与 Pawlak 的重要度是不同的。在广义粗糙集理论中，如果属性的广义重要度为零，则该属性是可以 δ 约简的。那么，可以在最初的计算单个属性的广义重要度时，使一部分属性得到约简。

文中采用基于属性的广义重要度的 Beam 搜索方法进行属性的 δ 近似约简[10]。

输入：决策系统 $S = \langle U, C \cup d \rangle$，$U/d = \{D_1, D_2, \cdots, D_{r(d)}\}$，$C = \{a_1, a_2, \cdots, a_m\}$；

输出：属性集 C 的多个近似约简集合。

算法步骤如下：

1) 计算 $\sigma_g(a_i)(i = 1, 2, \cdots, m)$，形成 m 个1-元组；计算 $\sigma_g(C)$。

2) $C = C \backslash a_j(\sigma_g(a_j) = 0)$。

3) 从属性集 C 中选择 k 个最大广义属性重要度值的属性（k 表示 Beam 的宽度），以这些属性为起点开始属性子集的搜索；

4) 加一个新属性到 k 个属性中的每一个上面，形成 $k(|C| - 1)$ 个2-元组，此时元组的尺寸为 $t = 2$；

5) 计算每一个 t-元组的属性子集的广义重要度，并选择广义重要度最大的 k 个属性子集；

6) 在这 k 个 t-元组的基础上加上一个新属性形成所有可能的 $(t+1)$-元组；

7) 重复5）到6），直到某个 t-元组形成的属性子集的广义重要度的值与 $\sigma_g(C)$ 的关系满足定义5，那么这个 t-元组搜索停止，而其他的 $(k-1)$ 个 t-元组则继续搜索，直到也能得到它的属性子集的广义重要度的值近似等于 $\sigma_g(C)$ 则停止；

8) 最后得到的与 $\sigma_g(C)$ 相等的 t-元组对应的属性子集就是属性的近似约简集合。

该约简算法的时间复杂度为 $O(km \cdot r(d)\log(m) \cdot C_{r(d)}^2)$，空间复杂度为 $O(2kmC_{r(d)}^2)$。因为算法只与属性集合的数目和决策类别数有关，而与样本数无关，因此对于大的样本集合，算法的时间较短。

定义6 给定决策系统为 $S = \langle U, C \cup d \rangle$，$U/d = \{D_1, D_2, \cdots, D_{r(d)}\}$。设 $B_1, B_2 \in RED^\delta(C)$，都是属性 C 的 δ 约简，如果 $\sum_{i=1}^{r(d)} |BN_{B_1}^\delta(d_i)| < \sum_{i=1}^{r(d)} |BN_{B_2}^\delta(d_i)|$，则称约简 B_1 的质量好于约简 B_2。

定义5的 δ 属性约简考察的是约简的属性与原属性集合有同样的全局分类能力，没有考虑到对确定性决策分类，还是可能性分类的具体影响。但是实际解决问题时，往往希望得到确定的结果，或希望不确定的结果越少越好。因此，采用定义6的方法，用各个约简所能产生的不确定分类的数目大小评价约简集合的质量优劣，是实际可行的。

4 实例验证

为了验证本文提出方法的有效性，文中采用经典粗糙集理论和相容粗糙集理论的分类方法与之相比较。实验所需要的数据集是从 UCI Machine Learning Repository 下载的 Iris 数据集和 Wine 数据集。每个数据集都被分成了两个部分：训练集（一半的数据）和测试集（整体数据）。

实验对比了三种分类方法的分类性能，结果如表3所示。当运用经典粗糙集理论进行分类时，笔者采用了 Rosetta 工具箱，其中选取 semi-naïve 算法进行离散化，选取遗传算法进行约简。

从表3仿真结果可以看出，后两种粗糙集方法的分类精度要明显好于第一种基于经典粗糙集理论的分类方法。而且，经典粗糙集理论的分类方法十分费时，其分类效果容易受

离散化过程的影响。相容粗糙集理论和本文的方法均可以直接处理实值属性,因此它们更为鲁棒。然而,相容粗糙集理论中的距离定义为属性值之间差值的绝对值,这种简单的定义可能会降低它用于分类时的精度。从实验的仿真结果看,广义粗糙集理论比经典粗糙集和相容粗糙集在分类问题上都更为有效。

表 3　三种不同粗糙集的分类性能比较

分类方法	Iris 数据集					Wine 数据集				
	约简	混合矩阵	被拒绝样本	错分类样本	精度/%	约简	混合矩阵	被拒绝样本	错分类样本	精度/%
经典粗糙集理论	a_2, a_3, a_4	$\begin{bmatrix}49&0&0\\0&47&2\\0&0&45\end{bmatrix}$	42th, 86th, 104th, 108th, 120th, 134th, 135th	78th, 84th	94.00	a_5, a_7, a_{10}, a_{11}	$\begin{bmatrix}52&0&46\\5&60&2\\0&0&48\end{bmatrix}$	42th, 86th, 104th, 108th, 120th, 134th, 135th	78th, 84th	94
相容粗糙集理论	—	$\begin{bmatrix}50&0&0\\0&48&1\\0&1&45\end{bmatrix}$	78th, 118th, 120th, 130th, 132th	71th, 150th	95.53	—	$\begin{bmatrix}59&0&0\\0&65&2\\0&0&48\end{bmatrix}$	42th, 86th, 104th, 108th, 120th, 134th, 135th	78th, 84th	94
δ 粗糙集理论	a_3, a_4	$\begin{bmatrix}50&0&0\\0&49&1\\0&4&46\end{bmatrix}$	None	71th, 107th, 120th, 134th, 135th	96.67	a_7, a_{10}, a_{12}, a_{13}	$\begin{bmatrix}58&1&0\\2&67&1\\0&0&48\end{bmatrix}$	42th, 86th, 104th, 108th, 120th, 134th, 135th	78th, 84th	94

5　结语

本文从实值决策表中直接分析各个属性的广义重要度,并利用该重要性进行空间距离的加权度量,提高了样本之间的可分性。另外,文中主要研究了实数域中粗糙集理论的一些问题,如 δ 广义粗糙下、上近似集,实值 δ 属性约简等,并对多个属性约简结果进行了约简质量的分析。最后,通过对两个分类问题的实例仿真验证了广义粗糙集理论是有效的。

参考文献:

[1] HAN JIAN-CHAO, SANCHEZ R, HU XIAO-HUA. Feature selection based on relative attribute dependency: An experimental study [C]// Proceedings of International Conference on Rough Set, Fuzzy Set, Data Mining and Granular Computing, LNCS 3641. Berlin: Springer-Verlag, 2005: 214 - 223.

[2] SWINIARSKI R W, SKOWRON A. Rough set methods in feature selection and recognition [J]. Pattern Recognition Letters, 2003, 24 (6): 833 - 849.

[3] SU C-T, HSU J-H. An extended Chi2 algorithm for discretization of real value attributes [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 437 - 441.

[4] NGUYEN S H, SKOWRON A. Quantization of real value attributes: Rough set and Boolean reasoning approach[J]. Bulletin of International Rough Set Society, 1996(1): 5 - 16.

[5] ROY A, PAL S K. Fuzzy discretization of feature space for a rough set classifier [J]. Pattern Recognition Letters, 2003, 24(6): 895 - 902.

[6] SLEZAK D, ZIARKO W. Variable precision Bayesian rough set model [C]// Proceedings of 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing (RSFDGrC), LNAI 2639. Berlin: Springer-Verlag, 2003: 312 - 315.

[7] SKOWRON A, STEPANIUK J. Tolerance approximation spaces [J]. Fundamenta Informaticae, 1996, 27(2/3): 245 - 253.

[8] HO T B, KAWASAKI S, NGUYEN N B. Cluster-based information retrieval with a tolerance rough set model [J]. International Journal of Fuzzy Logic and Intelligent Systems, 2002, 2(1): 26 - 32.

[9] NGUYEN H S, SLEZAK D. Approximate reducts and association rules-correspondence and complexity results [C]// Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, LNCS 1711. Berlin: Springer-Verlag, 1999: 137 - 145.

[10] GUPTA P, DOERMANN D, DEMENTHON D. Beam search for feature selection in automatic SVM defect classification [J]. Pattern Recognition, 2002, 19(2): 212 - 215.

[11] DAIJIN K. Data classification based on tolerant rough set [J]. Pattern Recognition, 2001, 34(8): 1613 - 1624.

(上接第 1419 页)

天数 > 0.37,当月停机次数 > 1,停机天数 > 14.4,3 月平均出账费用(包括欠费费用) > 126.95 元→流失(可信度为 100%)。

结论　这类用户当月费用较低,且停机天数较多,或者是停机天数较多,当月费用较大(存在恶意欠费嫌疑)的用户。

4　结语

在实际应用中,造成客户流失的原因比较多,因此不同的原因产生的客户流失判断模式也不一样,传统的客户流失只针对客户流失群体进行整体的模式判断,从而导致预测结果的精度比较低。本文针对实际应用中的客户流失样本分布多样性的特点,提出了一种基于多模式分的分类算法。通过与 Logistic、决策树、神经网络等方法的实践应用结果表明,新算法在客户流失预测精度上得到了较大的提高。

参考文献:

[1] LIU TSUNG-CHI, WU LI-WEI. Customer retention and cross-buying in the banking industry: An integration of service attributes, satisfaction and trust [J]. Journal of Financial Services Marketing, 2007, 12(2): 132 - 145.

[2] 王雷,陈松林,顾学道. 客户流失预警模型及其在电信企业的应用[J]. 电信科学, 2006, 22(9): 47 - 51.

[3] 郭明,郑惠莉,卢毓伟. 基于贝叶斯网络的客户流失分析[J]. 南京邮电学院学报, 2005, 25(5): 79 - 83.

[4] 叶进,林士敏. 基于贝叶斯网络的推理在移动客户流失分析中的应用[J]. 计算机应用, 2005, 25(3): 673 - 675.

[5] 桂现才,彭宏,王小华. C4.5 算法在保险客户流失分析中的应用[J]. 计算机工程与应用, 2005, 41(17): 197 - 199.

[6] 钱苏丽,何建敏,王纯麟. 基于改进支持向量机的电信客户流失预测模型[J]. 管理科学, 2007, 20(1): 54 - 58.