基于容错改进的邻域粗糙集属性约简算法*

彭潇然[a], 刘遵仁[b], 纪 俊[b]
(青岛大学 a. 数据科学与软件工程学院; b. 计算机科学技术学院, 山东 青岛 266071)

摘 要: 作为 Pawlak 粗糙集的扩展,邻域粗糙集能有效地处理数值型的数据。但是,因为沿用了 Pawlak 粗糙集在构造上下近似集时的包含关系,邻域粗糙集对噪声数据的容错性很差。针对这个问题,通过引入贝叶斯最小风险决策规则,提出了一种基于容错改进的邻域粗糙集属性算法。通过与现有的算法进行比较,实验结果表明,在数据预处理阶段用该算法能得到更好的属性约简。

关键词: 粗糙集; 邻域粗糙集; 决策粗糙集; 属性约简; 容错性
中图分类号: TP18 **文献标志码**: A **文章编号**: 1001-3695(2018)08-2256-04
doi:10.3969/j.issn.1001-3695.2018.08.004

Attribute reduction algorithm based on fault-tolerance improvement of neighborhood rough set

Peng Xiaoran[a], Liu Zunren[b], Ji Jun[b]
(*a. College of Data Science & Software Engineering, b. College of Computer Science & Technology, Qingdao University, Qingdao Shandong* 266071, *China*)

Abstract: As the extension of Pawlak rough set, neighborhood rough set can effectively deal with numerical data. However, its fault tolerance is very poor to noise data, because it follows the inclusion relation which is used for constructing the upper and lower approximations in Pawlak rough set. In order to solve this problem, this paper presented a new algorithm based on fault-tolerance improvement of neighborhood rough set by introducing the Bayes decision with minimum risk. Compared with the existing algorithm, the experimental results show that the attribute reduction obtained by this proposed algorithm is better in the data pre-processing.

Key words: rough set; neighborhood rough set; decision rough set; attribute reduction; fault tolerance

粗糙集理论认为知识是有粒度的,它是一种对论域中对象进行分类的能力。经典的 Pawlak 粗糙集[1]引入等价关系,将论域划分为多个等价类(信息粒),然后根据等价类与决策类之间的包含关系提出了上下近似的概念。通过运用上下近似的概念,Pawlak 粗糙集能有效地处理模糊和不精确的问题,但是在实际应用中也遇到了一些问题。

为了解决等价关系只适用于处理离散型数据,而不适用于处理数值型数据的问题,Zadeh[2]提出了信息粒化和粒度计算的概念。Lin[3]在信息粒化、粒度的基础上提出了邻域模型的概念。Hu 等人[4]提出了基于邻域粒化和粗糙逼近的决策表属性约简算法。作为 Pawlak 粗糙集的扩展,邻域粗糙集[2~6]可以有效地处理离散型和数值型的数据。但是,邻域粗糙集沿用了 Pawlak 粗糙集在构造上下近似时的包含关系,这使得邻域粗糙集在解决实际问题时缺乏一定的容错能力。

为了解决单纯的包含关系造成的零容错问题,研究人员在 Pawlak 粗糙集的基础上提出了多种概率粗糙集模型,例如 0.5-概率粗糙集模型[7]、变精度粗糙集模型[8]、贝叶斯粗糙集模型[9,10]等。其中,决策粗糙集模型(decision-theoretic rough set, DTRS)[11,12]引入了最小风险决策规则,在处理实际问题时具有更好的容错性,因而得到了广泛的关注和研究。但是,决策粗糙集沿用了 Pawlak 粗糙集在进行知识划分时的等价关系,这使得决策粗糙集不能直接处理数值型的数据。

基于以上研究,针对邻域粗糙集模型中存在的零容错问题,本文通过引入并重新定义决策粗糙集模型中的最小风险决策规则,进而提出了一种具有一定容错性的邻域粗糙集属性约简算法(attribute reduction algorithm based on fault-tolerance improvement, ARABFTI)。在 ARABFTI 算法中,将最小风险决策规则推导至 $\Pr(X|\delta(x_i))$ 与阈值 α、β 之间的比较,并且在邻域空间下对 $\Pr(X|\delta(x_i))$ 进行了新的定义。在进行正域计算时,会依据 $\Pr(X|\delta(x_i))$ 与阈值 α 的大小关系对对象进行决策,从而改进了领域粗糙集的容错性。

1 邻域粗糙集

1.1 邻域粒化

相较 Pawlak 粗糙集,邻域粗糙集中信息粒的并集与论域由等价关系变成了覆盖关系。

定义 1 给定 n 维实数空间 R^n,对于空间中的任意两个点 $x_i=(x_{i1},x_{i2},\cdots,x_{in})$ 和 $x_j=(x_{j1},x_{j2},\cdots,x_{jn})$,定义 $d(x_i,x_j)$ 是 R^n 上的一个度量计算,满足

$$d(x_i,x_j)=\left(\sum_{p=1}^{n}|x_{ip}-x_{jp}|^2\right)^{\frac{1}{2}}$$

定义 2 在实数空间上,定义对象的非空有限集合 $U=\{x_1,x_2,\cdots,x_n\}$,且称 U 为论域。定义 U 上的任意对象 x_i 的 δ-邻域为 $\delta(x_i)=\{x_j|x_j\in U,d(x_i,x_j)\leqslant\delta\}$,其中 $\delta\geqslant 0$。$\delta(x_i)$ 称做由 x_i 生成的 δ-邻域信息粒子,简称为 x_i 的邻域粒子。

1.2 邻域决策系统

定义 3 四元组 NDT = $(U,C\cup D,V,f)$ 为邻域决策系统(决策表)。其中 U 是论域;C 是条件属性集;D 是决策属性集,且 $C\cap D=\varnothing$,$C\neq\varnothing$,$D\neq\varnothing$;V 是信息函数 f 的值域。

定义 4 对于一个给定的邻域决策系统 NDT = $(U,C\cup D,V,f)$,D 将 U 划分为 N 个决策(等价)类,即 $D_1,D_2,\cdots,D_N$,$\forall B\in C$,定义决策属性集 D 关于 B 的下近似和上近似为

$$\underline{N_BD}=\bigcup_{i=1}^{N}\underline{N_BD_i},\overline{N_BD}=\bigcup_{i=1}^{N}\overline{N_BD_i}$$

收稿日期: 2017-04-12; **修回日期**: 2017-05-12 **基金项目**: 国家自然科学基金资助项目(61503208)

作者简介: 彭潇然(1994-),男,湖北天门人,硕士,主要研究方向为粗糙集理论(pxrl203@qq.com);刘遵仁(1963-),男,副教授,博士,主要研究方向为粗糙集理论、数据挖掘、智能计算等;纪俊(1982-),男,副教授,博士,主要研究方向为数据挖掘、大数据技术、转化医学等.

其中:$\underline{N_B D_i}=\{x_i\mid\delta_B(x_i)\subseteq D_i,x_i\in U\}$,$\overline{N_B D_i}=\{x_i\mid\delta_B(x_i)\cap D_i\neq\varnothing,x_i\in U\}$。

根据定义 1:

$$\delta_B(x_i)=\{x\mid d(B(x_i),B(x))\leqslant\delta,x\in U\}$$

定义决策属性集 D 关于 B 的正域为 $\mathrm{POS}_B(D)=\underline{N_B D}$,边界域为 $\mathrm{BND}_B(D)=\overline{N_B D}-\underline{N_B D}$,负域为 $\mathrm{NEG}_B(D)=U-\overline{N_B D}$。

2　邻域空间下的决策风险分析

2.1　最小风险决策规则

根据决策粗糙集[11,12],本节将对在领域空间下形成邻域粒子 $\delta(x_i)$ 的 x_i 进行决策规则分析。

设对于对象 x_i 有两种互补的状态,x_i 属于或不属于某个决策类 X,且有以下三种决策:a_p,将 x_i 划入 X 的正域;a_B,将 x_i 划入 X 的边界域;a_N,将 x_i 划入 X 的负域。各决策的风险如表 1 所示。

表 1　决策风险函数

决策	X(正例)	X^C(负例)
a_p	$\lambda_{PP}=\lambda(a_P\mid X)$	$\lambda_{PN}=\lambda(a_P\mid X^C)$
a_B	$\lambda_{BP}=\lambda(a_B\mid X)$	$\lambda_{BN}=\lambda(a_B\mid X^C)$
a_N	$\lambda_{NP}=\lambda(a_N\mid X)$	$\lambda_{NN}=\lambda(a_N\mid X^C)$

在表 1 中,$\lambda_{PP}=\lambda(a_P\mid X)$ 表示在 x_i 属于 X 的情况下,将 x_i 划入 X 的正域所产生的风险;λ_{BP}、λ_{NP}、λ_{PN}、λ_{BN}、λ_{NN} 以此类推。所以在一般情况下,以上各风险满足关系:$\lambda_{PP}\leqslant\lambda_{BP}\leqslant\lambda_{NP}$,$\lambda_{PN}\geqslant\lambda_{BN}\geqslant\lambda_{NN}$。即正例错分的风险逐级递增,负例正分的风险逐级递减。根据表 1,进一步得到对 x_i 采取不同的决策所产生的风险,表示如下:

$$R(a_P\mid x_i)=\lambda_{PP}\Pr(X\mid\delta(x_i))+\lambda_{PN}\Pr(X^C\mid\delta(x_i))$$
$$R(a_B\mid x_i)=\lambda_{BP}\Pr(X\mid\delta(x_i))+\lambda_{BN}\Pr(X^C\mid\delta(x_i))$$
$$R(a_N\mid x_i)=\lambda_{NP}\Pr(X\mid\delta(x_i))+\lambda_{NN}\Pr(X^C\mid\delta(x_i))$$

其中:$\Pr(X\mid\delta(x_i))$ 表示在形成邻域粒子 $\delta(x_i)$ 的情况下,x_i 属于 X 的概率。根据最小风险决策规则:

a)若 $R(a_P\mid x_i)\leqslant R(a_B\mid x_i)$ 且 $R(a_P\mid x_i)\leqslant R(a_N\mid x_i)$,则 $x_i\in\mathrm{POS}(X)$。

b)若 $R(a_B\mid x_i)\leqslant R(a_P\mid x_i)$ 且 $R(a_B\mid x_i)\leqslant R(a_N\mid x_i)$,则 $x_i\in\mathrm{BND}(X)$。

c)若 $R(a_N\mid x_i)\leqslant R(a_P\mid x_i)$ 且 $R(a_N\mid x_i)\leqslant R(a_B\mid x_i)$,则 $x_i\in\mathrm{NEG}(X)$。

因为 $\Pr(X\mid\delta(x_i))+\Pr(X^C\mid\delta(x_i))=1$,用 $\Pr(X\mid\delta(x_i))$ 进一步简化上述规则,整理可以得到三个阈值。

$$\alpha=\frac{\lambda_{PN}-\lambda_{BN}}{(\lambda_{PN}-\lambda_{BN})+(\lambda_{BP}-\lambda_{PP})}$$
$$\gamma=\frac{\lambda_{PN}-\lambda_{NN}}{(\lambda_{PN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$
$$\beta=\frac{(\lambda_{BN}-\lambda_{NN})}{(\lambda_{BN}-\lambda_{NN})+(\lambda_{NP}-\lambda_{BP})}$$

可以证明 $\alpha\in(0,1]$,$\gamma\in(0,1)$,$\beta\in[0,1)$,上述规则重新表示为

a)若 $\Pr(X\mid\delta(x_i))\geqslant\alpha$ 且 $\Pr(X\mid\delta(x_i))\geqslant\gamma$,则 $x_i\in\mathrm{POS}(X)$。

b)若 $\Pr(X\mid\delta(x_i))\leqslant\alpha$ 且 $\Pr(X\mid\delta(x_i))\geqslant\beta$,则 $x_i\in\mathrm{BND}(X)$。

c)若 $\Pr(X\mid\delta(x_i))\leqslant\beta$ 且 $\Pr(X\mid\delta(x_i))\leqslant\gamma$,则 $x_i\in\mathrm{NEG}(X)$。

当且仅当 $\dfrac{\lambda_{NP}-\lambda_{BP}}{\lambda_{NP}-\lambda_{NN}}>\dfrac{\lambda_{BP}-\lambda_{PP}}{\lambda_{PN}-\lambda_{BN}}$,可得 $\alpha>\gamma>\beta$,此时上述规则进一步简化为:

a)若 $\Pr(X\mid\delta(x_i))\geqslant\alpha$,则 $x_i\in\mathrm{POS}(X)$。

b)若 $\Pr(X\mid\delta(x_i))<\alpha$ 且 $\Pr(X\mid\delta(x_i))>\beta$,则 $x_i\in\mathrm{BND}(X)$。

c)若 $\Pr(X\mid\delta(x_i))\leqslant\beta$,则 $x_i\in\mathrm{NEG}(X)$。

2.2　$\Pr(X\mid\delta(x_i))$ 的定义及容错性分析

在不同的决策模型中,对 $\Pr(X\mid\cdot)$($\cdot$ 表示在不同模型中不同的分析对象)的定义不同[13,14]。本文提出一种邻域空间下的 $\Pr(X\mid\cdot)$:

定义 5　$\Pr(X\mid\delta(x_i))=\dfrac{\lvert\{x\mid x\in\delta(x_i)\&D(x)=D(x_i)\}\rvert}{\lvert\delta(x_i)\rvert}$

即 x_i 属于 X 的概率等于邻域粒子 $\delta(x_i)$ 中和 x_i 类别相同的对象个数比上对象总个数的比值。

在该定义下,基于图 1 中的对象分布,本文将对比引入风险决策规则前后的邻域粗糙集对噪声数据的容错性。

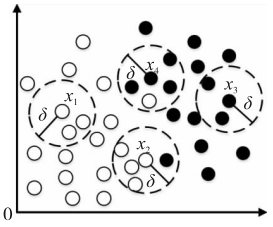


图1　具有容错性的情况

根据定义 4,在经典的邻域粗糙集中,$\mathrm{POS}_B(D_i)=\{x_i\mid\delta_B(x_i)\subseteq D_i,x_i\in U\}$,分析图 1 中的四个标注对象:空心圆和实心圆分别表示对象的两种类别,用 D_W 表示空心圆等价类,D_B 表示实心圆等价类,虚线圆圈表示某个对象的 $\delta(x_i)$ 邻域。可知 $\delta(x_1)\subseteq D_W$,$\delta(x_2)\not\subset D_W$,$\delta(x_3)\subseteq D_B$,$\delta(x_4)\not\subset D_B$,即 $x_1,x_3\in\mathrm{POS}_B(D)$,$x_2,x_4\notin\mathrm{POS}_B(D)$。这种情况被认为:当前属性集 B 没有成功对对象 x_2 和 x_3 进行正确分类,需要增加新的属性进行进一步判别。但是,在 $\delta(x_2)$ 和 $\delta(x_4)$ 中各有五个对象,其中有四个对象的类别相同,仅有一个对象的类别和它们不同。在数据集中噪声数据存在的情况下,这说明当前属性集 B 已经能很好地对这两个对象进行分类了,此时若再增加属性,则增加的属性对当前属性集 B 的分类能力的贡献是无法准确判定的,这个属性不一定是重要属性,可能只是一般属性,甚至是冗余属性。这种对于噪声数据的零容错性,可以理解为

$$\Pr(X\mid\delta(x_i))=\begin{cases}0 & D(\exists x\in\delta(x_i))\neq D(x_i)\\ 1 & D(\forall x\in\delta(x_i))=D(x_i)\end{cases}$$

根据定义 5 和 2.1 节中的最小风险决策规则,设 $\alpha=0.8$,分析图 1 中的四个对象可知:$\Pr(X\mid\delta(x_1))=1\geqslant0.8$,$\Pr(X\mid\delta(x_2))=4/5\geqslant0.8$,$\Pr(X\mid\delta(x_3))=1\geqslant0.8$,$\Pr(X\mid\delta(x_4))=4/5\geqslant0.8$,即 $x_1,x_2,x_3,x_4\in\mathrm{POS}_B(D)$。可以看出,对对象 x_2 和 x_3 进行决策时避免了噪声数据对结果造成的影响。这说明在数据集中噪声数据存在的情况下,引入风险决策规则后,能在一定程度上提升邻域粗糙集的容错性,从而得到更好的属性约简。

3　基于容错改进的属性约简算法

3.1　基于正域计算的前向贪心约简算法

对一个决策表而言,如何设计有效的算法用于删除冗余属性,得到属性约简是粗糙集理论研究的重点之一。

定义 6[15]　对于一个决策表 $\mathrm{NDT}=(U,C\cup D,V,f)$,给定有限集合 $B\subseteq C$,若满足 $\mathrm{POS}_B(D)=\mathrm{POS}_C(D)$,则称 B 是一个独立属性子集;如果对 $\forall a\in B$,$\mathrm{POS}_{B-\{a\}}(D)<\mathrm{POS}_B(D)$,则称 B 为 C 的一个属性约简。

最小属性约简问题已被证明是一个 NP-hard 问题。为了在较短时间内得到一个较好的约简结果,基于正域计算,Hu 等人[4]采用贪心策略提出了 F2HARNRS(fast forward heterogeneous attribute reduction based on neighborhood rough sets)算法。F2HARNRS 算法的具体策略是:初始化属性约简集合为空集,此时约简集合下的正域为空集,每次选取使正域中对象增加最多的属性加入到约简集合中,直至正域中的对象不再增加,输出集合。其中,Hu 等人证明了 F2HARNRS 算法满足正域单调性(见定义 7),即约简集合中新增加的属性不会使已属于正域的样本变为非正域样本这一性质,在算法的计算过程中,每次仅对还未判定为正域的样本进行正域计算,进一步缩减了算法的时间开销。

F2HARNRS 算法提出后得到了广泛的应用与研究[16,17],是

邻域粗糙集算法中较被认可的算法。本文对邻域粗糙集算法容错性的研究也建立在对 F2HARNRS 算法的分析和对比之上。

3.2 本文的 ARABFTI 算法

3.2.1 ARABFTI 算法的正域计算

相较 F2HARNRS 算法的正域计算,ARABFTI 算法的正域计算重新定义了将对象划入正域的判断条件,且增大了算法计算量。

根据以上分析,提出 ARABFTI 算法的正域计算算法 POS(NDT,δ,α),如算法 1 所示。

算法 1

输入: NDT = (U,red$\cup D$,V,f),δ,α。
输出:正域 $\mathrm{POS}_{\mathrm{red}}(\mathrm{D})$。
a)初始化 $\mathrm{POS}_{\mathrm{red}}(D)=\varnothing$,$\delta(x_i)=\varnothing$
b)for each $x_i \in U$
  for each $x_j \in U$
    if $d(x_i,x_j)\leqslant\delta$
      $\delta(x_i)\leftarrow x_j$
    end if
  end for
  计算 $\Pr(D|\delta(x_i))$;
  if $\Pr(D|\delta(x_i))\geqslant\alpha$
    $\mathrm{POS}_{\mathrm{red}}(D)\leftarrow x_i$
  end if
end for
c)return $\mathrm{POS}_{\mathrm{red}}(D)$

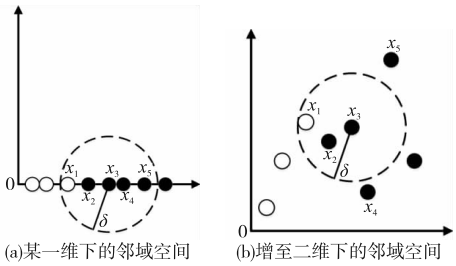
分析算法 1 中的 2 层循环计算:F2HARNRS 算法的正域计算因为算法的零容忍性,则在 2 层循环的计算过程中,若满足 $d(x_i,x_j)\leqslant\delta\&\&D(x_i)\neq D(x_j)$,可立即判定 $x_i\notin\mathrm{POS}_{\mathrm{red}}(D)$ 且立即跳出循环。而 ARABFTI 算法的正域计算由于需要计算 $\Pr(D|\delta(x_i))$,则会执行完 2 层循环,这会增加算法时间的开销。

3.2.2 ARABFTI 算法的正域单调性分析

相较 F2HARNRS 算法,ARABFTI 算法通过引入最小风险决策规则,具有了一定的容错性,但是也因此不满足正域单调性。

定义 7 对给定的决策表 NDT = (U,$C\cup D$,V,f),A 和 B 是 C 的两个子集,即 $A\subseteq C$,$B\subseteq C$。若对于 $A\subseteq B$,满足 $\mathrm{POS}_A(D)\subseteq\mathrm{POS}_B(D)$,则称满足正域单调性。

分析 ARABFTI 算法的正域单调性:在图 2 中,用虚线圆圈代表邻域粒子,设 $\alpha=0.8$,如图 2(a)所示,当前属性集集合为 $\{a\}$,则在一维的邻域空间中,$\delta(x_3)=\{x_1,x_2,x_3,x_4,x_5\}$,因为 $\Pr(D|\delta(x_3))=4/5\geqslant0.8$,所以 $x_3\in\mathrm{POS}_{\{a\}}(X)$;如图 2(b)所示,属性集集合增至 $\{a,b\}$,则在二维的邻域空间中,$\delta(x_3)=\{x_2,x_3,x_5\}$,因为 $\Pr(D|\delta(x_3))=2/3<0.8$,所以 $x_3\notin\mathrm{POS}_{\{a,b\}}(X)$,即 $\{a\}\subseteq\{a,b\}$,$\mathrm{POS}_{\{a\}}(X)\not\subset\mathrm{POS}_{\{a,b\}}(X)$,这说明 ARABFTI 算法不满足正域单调性。所以 F2HARNRS 算法中的部分思想不适用于 ARABFTI 算法。



(a)某一维下的邻域空间 (b)增至二维下的邻域空间

图2 不满足正域单调性的情况

3.2.3 ARABFTI 算法

ARABFTI 算法沿用了 F2HARNRS 的贪心思想,但是其不满足正域单调性,所以每次都需要对论域中的对象进行正域判断,这也会增加算法的时间开销。

根据以上分析,提出 ARABFTI 算法,如算法 2 所示。

算法 2

输入:NDT = (U,red,V,f),δ,α。
输出:red。
a)初始化 red = $\varnothing$, 当前正域 max_pos = $\varnothing$
使正域中对象增加最多的属性 max_i = $\varnothing$
原属性集下的正域 standard_pos = $\mathrm{POS}_C(D)$
b)while |max_pos| < |standard_pos|
  flag = 0;
  for each $k_i\in(C-\mathrm{red})$
  $\mathrm{pos}_i=\mathrm{POS}(\mathrm{NDT},\delta,\alpha)$;
    //NDT = (U,(red$\cup k_i$)$\cup D$,V,f)
    if |max_pos| < |pos_i|
      max_pos = Pos_i;
      max_i = k_i;
      flag = 1;
    end if
  end for
  if flag = 1
    red = red$\cup$max_i;
  else
    break;
  end if
end while
c)return red

比较两个算法正域计算的次数:假设某一数据集有 m 个属性,约简结果中包含 k 个属性,且每增加一个属性正域中增加 $|U|/k$ 个样本,则 F2HARNRS 算法进行正域计算的次数为

$$m|U|+(m-1)|U|\frac{k-1}{k}+\cdots+(m-k)|U|\frac{1}{k}<\frac{m|U|(1+2+\cdots+k)}{k}=\frac{m|U|(1+k)}{2}$$

ARABFTI 算法进行正域计算的次数为

$$m|U|+(m-1)|U|+\cdots+(m-k)|U|=\frac{(m+m-k)(m-(m+k)+1)}{2}=\frac{(2m-k)|U|(1+k)}{2}$$

由此可知,相较 F2HARNRS 算法,ARABFTI 算法牺牲了时间开销换取了一定程度的容错性。

4 实验分析

在本次实验中,首先分别用 F2HARNRS 算法和本文的 ARABFTI 算法在不同的 δ 取值下对数据集进行属性约简,比较属性约简个数;然后用 SVM 分类算法根据各自的属性约简对数据集进行分类,比较 SVM 算法的分类精度,精度越高代表属性约简越好。其中,在 2.1 节中提供了计算阈值 α、γ 和 β 的方法,可以根据实际的应用对风险参数进行设置,从而计算阈值[18]。在本次的实验分析中,取 $\alpha=0.6$ 和 $\alpha=0.8$,即在风险敏感度较低和较高这两种情况下进一步分析 F2HARNRS 算法。

4.1 实验环境及方案

UCI(University of California Irvine)(http://archive.ics.uci.edu/ml/)提供了一系列用于测试的标准数据集。本文从 UCI 中选取了三个属性数不同的数值型数据集,如表 2 所示。

表 2 数据集描述

编号	数据集	样本数	属性数	类别数
1	wine	178	13	3
2	WDBC	569	30	2
3	sonar	208	60	2

本次实验在一台 Intel® Pentium® CPU G620 和 4 GB 内存的 PC 机上,采用 Windows 7 环境下的 MATLAB R2016b 进行算法仿真。用 F2HARNRS 算法[4] 和 ARABFTI 算法分别对数据集进行约简。

根据定义 4 可知,δ 的取值直接影响着属性约简的结果。在不同的 δ 取值下,算法得到的属性约简不同,造成根据属性约简对数据集进行分类后,SVM 算法的分类精度不同。本文在[0.04,1]上,按 0.04 增进,共取得 25 个 δ 取值,记录在不同的 δ 取值下对应的属性约简个数和 SVM 算法的分类精度。其中,当 δ 增大至某个值时,得到的属性约简为空,称此时 δ 的值为最大取值点。

在 SVM 算法中,随机选取数据集中每类对象的 2/3 作为训练集,随机选取数据集中每类对象的 1/3 作为测试集,算法

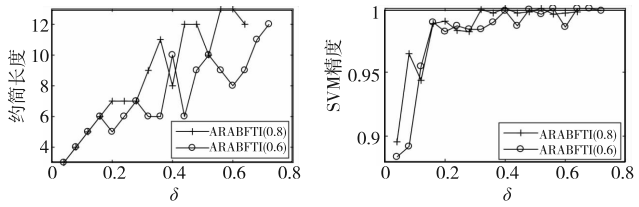
执行 20 次,分类精度的最后结果取均值。

4.2　F2HARNRS 和 ARABFTI 在各数据集上的对比

根据 2.2 节中的分析,在引入 $\Pr(X|\delta(x_i))$ 的前提下,F2HARNRS 算法的正域决策可以理解为 $\Pr(X|\delta(x_i))\geqslant 1$,即 $\alpha=1$。那么实验分析即是对 α 分别在 0.6、0.8 和 1 取值时的分析。为了更清晰直观地呈现实验结果,实验分为 ARABFTI(0.6)算法与 ARABFTI(0.8)算法、ARABFTI(0.8)算法与 F2HARNRS 算法之间的对比。

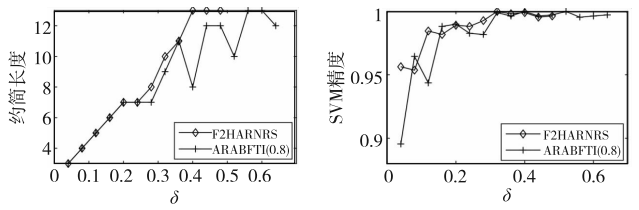
4.2.1　Wine 数据集

在 wine 数据集上的实验结果如图 3、4 所示。



(a)约简的属性个数　　(b)SVM算法的分类精度

图3　ARABFTI(0.6)和ARABFTI(0.8)在wine上的对比



(a)约简的属性个数　　(b)SVM算法的分类精度

图4　ARABFTI(0.8)和F2HARNRS在wine上的对比

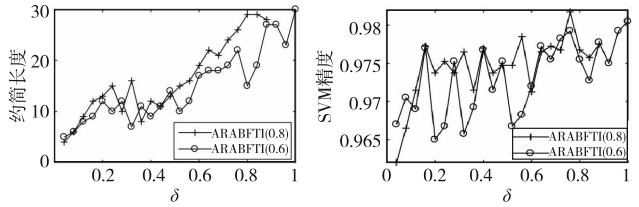
分析图 3(a)(b)可知,相较 ARABFTI (0.6)算法的实验结果,考虑到实验误差,ARABFTI(0.8)算法得到的属性约简个数较多且最大取值点较小,SVM 算法的分类精度较高。

分析图 4(a)(b)可知,相较 F2HARNRS 算法的实验结果,考虑到实验误差,ARABFTI(0.8)算法得到的属性个数差别不大且最大取值点较大,SVM 算法的分类精度大致相等。

以上实验结果说明,wine 数据集中噪声数据较少,相较 F2HARNRS 算法,ARABFTI 算法能保持所得属性约简的有效性。

4.2.2　WDBC 数据集

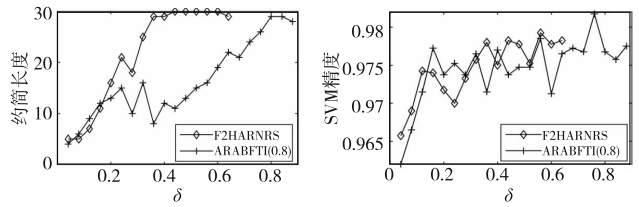
分析图 5(a)(b)能得到与图 3(a)(b)中一样的结论。



(a)约简的属性个数　　(b)SVM算法的分类精度

图5　ARABFTI(0.6)和ARABFTI(0.8)在WDBC上的对比

分析图 6(a)(b)可知,相较 F2HARNRS 算法的实验结果,考虑到实验误差,ARABFTI(0.8)算法得到的属性个数明显较少且最大取值点较大,SVM 算法的分类精度大致相等。



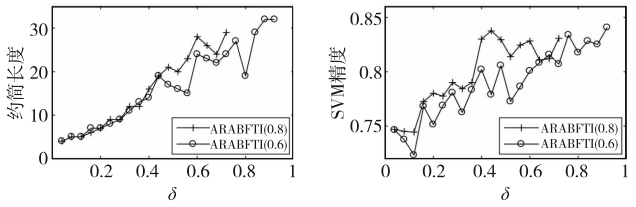
(a)约简的属性个数　　(b)SVM算法的分类精度

图6　ARABFTI(0.8)和F2HARNRS在WDBC上的对比

以上实验结果说明,WDBC 数据集中存在重要度很高的部分属性,仅根据这些属性就能对数据集进行可靠分类。相较 F2HARNRS 算法,ARABFTI 算法能在保持所得属性约简的有效性的基础上,得到属性个数更少的约简。

4.2.3　Sonar 数据集

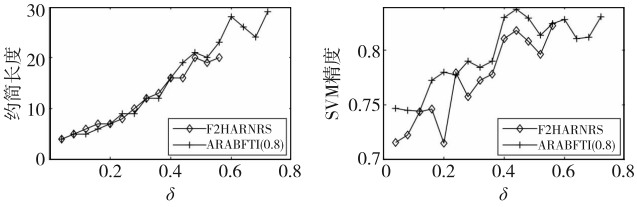
分析图 7(a)(b)能得到与图 3、5 中(a)(b)一样的结论。



(a)约简的属性个数　　(b)SVM算法的分类精度

图7　ARABFTI(0.6)和ARABFTI(0.8)在sonar上的对比

分析图 8(c)(d)可知,相较 F2HARNRS 算法的实验结果,考虑到实验误差,ARABFTI(0.8)算法得到的属性个数差别不大且最大取值点较大,SVM 算法的分类精度明显较高。



(a)约简的属性个数　　(b)SVM算法的分类精度

图8　ARABFTI(0.8)和F2HARNRS在sonar上的对比

以上实验结果说明,sonar 数据集中不存在重要度很高的部分属性,相较 F2HARNRS,ARABFTI 能在得到相同属性个数约简的基础上,得到有效性更高的约简,而这种有效性的提高是建立在具有一定容错能力的基础之上的。

4.3　实验结论

综上所述,在三个维数不同的数据集中,通过引入最小决策规则,ARABFTI 算法能在所得约简的属性个数和有效性上进一步提升 F2HARNRS 算法的性能。其中,风险敏感度越低,约简的属性个数越少,约简的有效性降低。在处理实际问题中,虽然阈值 α 由实际的风险参数计算得到,但是也不应过低。

5　结束语

相比零容错的 F2HARNRS 算法[4],本文的 ARABFTI 算法通过引入最小风险决策规则进一步提升了算法性能。与此同时,根据 3.2 节中的分析可知,这种容错性的提升带来了时间开销的增加,对比两种算法的正域计算次数的表达式进一步可知,这种时间开销上的增长是成倍的。如何让 ARABFTI 算法在具有一定容错性的前提上又具有较少的时间开销,这个问题笔者将在未来的工作中进行研究。

参考文献:

[1] Pawlak Z, So-Winski R. Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research, 1994,72(3):443-459.

[2] Zadeh L A. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets & Systems,1997,90(2):111-127.

[3] Lin T Y. Granular computing on binary relations I: data mining and neighborhood systems[J]. Rough Sets in Knowledge Discovery, 1998(2):165-166.

[4] Hu Qinghua, Yu Daren, Liu Jinfu, *et al*. Neighborhood rough set based heterogeneous feature subset selection[J]. Information Sciences,2008,178(18):3577-3594.

[5] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2001:147-156.

[6] 胡清华,于达人. 应用粗糙计算[M]. 北京:科学出版社,2012.

[7] Pawlak Z, Wong S K M, Ziarko W. Rough sets: probabilistic versus deterministic approach[J]. International Journal of Man-Machine Studies,1988,29(1):81-95.

(下转第 2314 页)

由表 4 可知,在步行、驾车以及公共交通这三种出行方式中,模型对于步行的识别率最高,查全率、查准率与 F 值均在 85% 以上,驾车与公共交通出行方式识别率达到了 80% 以上,整体判断准确率达到了 83.3%。

目前,部分学者利用单一的信令数据进行出行识别方面的研究,文献[12,13]中根据出行先验知识与模糊识别算法,利用不同出行方式构建出行平均速度、出行时长和出行距离等隶属度函数,从而进行出行方式的识别,但并未提供相应的识别结果。本文利用相同数据集,根据上述文献提供的识别方案与相关参数进行了实验并得出了结果:基于先验知识的出行方式模糊识别算法整体判断准确率为 68%,本文提出的基于手机信令和导航数据的居民出行方式识别方法精度提升超过 15%。两种方案的步行、驾车和公共交通出行识别具体结果对比如图 4 所示。由图 4 中可知,本文提出的识别方法比基于先验知识的出行方式模糊识别算法从查全率与查准率上均有所提升。

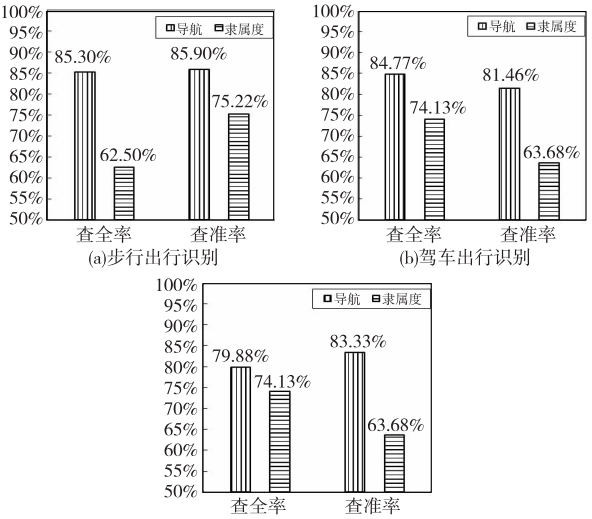


图4 两种方案识别结果的查全率、查准率对比

此外,算法模型执行效率是该算法模型能够在实际生产环境中部署的重要因素之一。本文算法模型的主要开发平台为 Eclipse,版本为 4.4;网络带宽为 4 Mbps,无其他程序占用网络。经过对相同数据集的 10 次测算,该模型的平均运行时间为 187 s,除去程序初始化以及读取文件等时间开销,平均运行时间为 173 s,识别速度较快。但该模型由于需要使用百度地图 API,所以需要一定的网络带宽来保障识别速度。

整体结果表明,本文提出的出行方式识别方法可以从轨迹相似性以及时间相似性等多角度对出行方式进行更加准确的识别,并且算法执行效率高,可用于工程环境中对居民出行方式进行识别。

3 结束语

居民的出行方式是城市交通规划的一个重要依据。目前,随着智能终端设备的普及,手机信令数据也成为交通研究中一个重要的数据源。本文基于手机信令数据与导航数据,利用 DBSCAN 聚类算法进行轨迹聚类,同时进行时间关联,设计了一种识别居民出行方式的新方法。在基于 502 个居民某天某时段的手机信令数据和导航数据的基础上,利用该方法得到的结果整体正确率达到了 83.3%,高于单一使用信令数据进行出行方式识别的方案,表明该方法可以更为准确地对居民的出行方式进行识别。此外,在本文的基础上,可以更加关注于多源数据的融合,如人物画像、公交 IC 卡和实时路况指数等数据,进一步提升居民出行方式的识别正确率。

参考文献:

[1] 冉斌. 手机数据在交通调查和交通规划中的应用[J]. 城市交通, 2013, 11(1): 72-81.

[2] Jahangiri A, Rakha H A. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data [J]. IEEE Trans on Intelligent Transportation Systems, 2015, 16(5): 2406-2417.

[3] 李喆,孙健,倪训友. 基于智能手机大数据的交通出行方式识别研究[J]. 计算机应用研究, 2016, 33(12):3527-3529,3558.

[4] Shafique M A, Hato E. Use of acceleration data for transportation mode prediction [J]. Transportation, 2015, 42(1): 163-188.

[5] Xu Chao, Ji Minhe, Chen Wen, *et al*. Identifying travel mode from GPS trajectories through fuzzy pattern recognition [C]//Proc of the 7th International Conference on Fuzzy Systems and Knowledge Discovery. Piscataway, NJ: IEEE Press, 2010: 889-893.

[6] Abdelaziz A M, Youssef M. The diversity and scale matter: ubiquitous transportation mode detection using single cell tower information [C]//Proc of the 81st IEEE Vehicular Technology Conference. Piscataway, NJ: IEEE Press, 2015: 1-5.

[7] Luo Jiangtao, Shu Zhonglin, Zhou Yunfeng, *et al*. Monitoring system of urban population traffic based on mobile network signaling[C]// Proc of IEEE/CIC International Conference on Communications in China. Piscataway, NJ: IEEE Press, 2014: 339-343.

[8] 赖见辉, 陈艳艳, 钟园, 等. 基于手机定位信息的地铁乘客出行路径辨识方法[J]. 计算机应用, 2013, 33(2): 583-586.

[9] Bloch A, Erdin R, Meyer S, *et al*. Battery-efficient transportation mode detection on mobile devices[C]// Proc of the 16th IEEE International Conference on Mobile Data Management. Piscataway, NJ: IEEE Press, 2015: 185-190.

[10] Wang Huayong, Calabrese F, Lorenzo G D, *et al*. Transportation mode inference from anonymized and aggregated mobile phone call detail records[C]// Proc of the 13th IEEE International Conference on Intelligent Transportation Systems. Piscataway, NJ: IEEE Press, 2010: 318-323.

[11] Ester M, Kriegel H P, Sander J, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proc of International Conference on Knowledge Discovery and Data Mining. Palo Alto, CA: AAAI Press, 1996: 226-231.

[12] 张博. 基于手机网络定位的 OD 调查的出行方式划分研究[D]. 北京: 北京交通大学, 2010.

[13] 冯冲. 基于移动定位数据的用户出行模式识别[D]. 昆明: 昆明理工大学, 2011.

(上接第 2259 页)

[8] Ziarko W. Variable precision rough set model[J]. Journal of Computer & System Sciences, 1993, 46(1): 39-59.

[9] Zak D. Rough sets and Bayes factor[M]. Berlin: Springer-Verlag, 2005: 53-63.

[10] Lezak D, Ziarko W. The investigation of the Bayesian rough set model [J]. International Journal of Approximate Reasoning, 2005, 40(1): 81 -91.

[11] Yao Yiyu. Decision-theoretic rough set models[C]//Lecture Notes in Computer Science, vol 4481. Berlin: Springer, 2007: 1-12.

[12] Yao Y Y, Wong S K M. A decision theoretic framework for approximating concepts[J]. International Journal of Man-Machine Studies, 1992, 37(6): 793-809.

[13] Yao Yiyu, Zhou Bing. Naive Bayesian rough sets[C]//Proc of International Conference on Rough Set & Knowledge Technology. Berlin: Springer, 2010: 719-726.

[14] 郭敏, 贾修一, 商琳. 基于模糊化的决策粗糙集属性约简和分类[J]. 模式识别与人工智能, 2014, 27(8): 701-707.

[15] Pawlak Z, Slowinski R. Rough set approach to multi-attribute decision analysis[J]. European Journal of Operational Research, 1994, 72(3): 443-459.

[16] Liu Yong, Huang Wenliang, Jiang Yunliang, *et al*. Quick attribute reduct algorithm for neighborhood rough set model[J]. Information Sciences, 2014, 271(7): 65-81.

[17] 刘遵仁, 吴耿锋. 基于邻域粗糙模型的高维数据集快速约简算法[J]. 计算机科学, 2012, 39(10): 268-271.

[18] 于洪, 王国胤, 姚一豫. 决策粗糙集理论研究现状与展望[J]. 计算机学报, 2015, 38(8): 1628-1639.