

基于 GA 与 RST 的分类规则挖掘算法

谢娟英¹ 刘芳² 冯德民¹

(陕西师范大学计算机科学学院 西安 710062)¹ (西安电子科技大学计算机学院 西安 710071)²

摘要 本文提出了在没有任何领域知识可供借鉴的情况下,利用遗传算法对信息系统的数量型属性进行离散化,利用 RST 进行分类规则挖掘,将 GA 与 RST 相结合进行分类规则挖掘的新算法。该算法不仅有效地解决了利用粗糙集理论进行分类规则挖掘时,数量型属性的离散化问题,而且可挖掘出通用的分类规则。

关键词 遗传算法,离散化,粗糙集理论,分类规则挖掘

Mining Classification Rules Based on GA and RST

XIE Juan-Ying¹ LIU Fang² Feng De-Min¹

(College of Computer Science, Shanxi Normal University, Xi'an 710062)¹

(School of Computer Science and Engineering, Xidian University, Xi'an 710071)²

Abstract A new algorithm for mining classification rules based on GA and RST is presented in this paper. And it is implemented in VC++ 6.0. The algorithm uses GA to discretizing values of quantitative attributes of an information system without any experienced knowledge about a special field, and uses RST for mining general classification rules. Its validity is proved by examples.

Keywords Genetic algorithm, Discretization, Rough sets theory, Mining classification rules

1 引言

分类规则挖掘就是通过对训练集数据的分析,产生关于分类的规则或精确描述^[1],以用于对未知类别的数据进行分类预测。在医学诊断、经济及商业预测、卫星信号分类分析^[2~5]等领域有着广泛的应用。

粗糙集理论(Rough Sets Theory, RST)是对不精确、模糊、不确定数据进行分析、学习的主要方法^[6]。该方法只依赖于数据内部的知识,用数据之间的近似来表示知识的不确定性^[7],因此很适合用来做分类规则的挖掘。应用粗糙集进行分类规则挖掘,需要对样本属性进行编码,这要求对样本的连续型属性先进行离散化处理。

数量型(连续型)属性离散化,就是要给数量型属性找到合适的分段点,这可看作一个聚类问题。对于有领域知识可供借鉴的问题,样本数量属性的离散化可依据领域知识直接进行,领域知识不足时可用传统的 C 均值聚类算法^[8]。但是 C 均值算法受初始聚类中心的影响较大,对类间明确的样本分类效果较好,而在类间模糊时,分类效果欠佳。

遗传算法(Genetic Algorithm, GA)使用群体搜索技术,通过对当前群体施加选择、交叉、变异等一系列遗传操作算子,产生出新一代的群体,并逐步使群体进化到包含或接近最优解的状态^[9]。因此,在没有任何领域知识的情况下,可利用它对样本的数量型属性进行离散化,以找到连续型属性的最佳分段点。

本文利用遗传算法具有全局寻优能力的特性,以及 RST 理论处理数据的优点,提出了 GA 与 RST 理论相结合的分类规则挖掘方法,该方法可挖掘出通用的分类规则。

2 GA 与 RST 相结合的分类规则挖掘算法

该算法包括两部分:一是基于 GA 的数量型属性离散化

方法;二是基于 RST 的分类规则提取方法。

2.1 基于 GA 的属性离散化方法

算法思想:利用遗传算法本身的特性,随机产生初始聚类(个体),然后利用相应的遗传操作算子,使聚类向最佳聚类(最佳个体)逼近,当满足一定的结束条件时退出。此时的最佳个体即为最佳聚类。

为了找到样本每一个数量型属性合适的分段点,分别对样本根据其每一个数量型属性进行聚类。根据聚类结果,对每一个类的样本属性按值排序,选择两类间相邻属性值的平均值作为分段点。为了确定每一个属性的分类数,让 C(聚类数)从 1 变到 M(训练样本数),根据 J_e 随 C 的变化曲线确定合适的 C 值^[8]。

算法实现:个体的编码,采用 0、1 编码,长度为 $C * M$,其中, C 为聚类数, M 为训练样本数,采用随机选择方法,若某一个样本被选中,则相应位编 1,否则编 0。由于个体编码的特点,遗传操作,仅包含选择和变异操作。选择为轮盘赌随机选择,变异概率 $P_m = 0.1$ 。种群大小为 80,最大进化代数 200。个体适应度函数定义为聚类误差平方和 J_e 与某一非零常量的倒数,即:

$$f_{fitness} = \frac{1}{J_e + K}$$

2.2 基于 RST 的分类规则挖掘方法

利用 2.1 节提出的基于 GA 的属性离散化方法为样本的每一个连续型条件属性求得最佳的分段点,再对不同段赋以不同的编码。经过该数据预处理原信息系统(决策表)转化为属性值是离散型的决策表。其中,信息系统(Information System, IS)为: $IS = (U, A, V, f)$, $U = \{x_1, x_2, x_3, x_m\}$, $A = C \cup D$, $C \cap D = \emptyset$, C 是条件属性集, D 是决策属性集, $card(D) \geq 1$, $V = \bigcup_{a \in A} V_a$ 是属性值的集合, V_a 是属性 a 的值域, $f: U \times A$

→ V 是一个信息函数,对每一个 $a \in A$, 和 $x \in U$, $f(x, a) \in V_a^{[6 \sim 7, 10 \sim 12]}$ 。

基于 RST 的分类规则提取步骤如下:

step1: 分别计算决策空间 D 的基本集 U/D 及条件属性空间 C 的基本集 U/C ;

step2: 计算 D 基本集相对于 C 基本集的上下近似 $\underline{C}(U/D)$, $\overline{C}(U/D)$, 并计算规则精度;

step3: 构造 C 的 D -可辨识矩阵及可辨识函数 $f(C)$ 。其中 C 的 D -可辨识矩阵是 $n \times n$ 对称矩阵, $n = \text{card}(U)$, 其任一元素 $\alpha(x, y) = \{a \in A \mid f(x, a) \neq f(y, a), x \in [x]_{\text{Ind}(D)}, y \in [y]_{\text{Ind}(D)} \text{ 且 } [x]_{\text{Ind}(D)} \neq [y]_{\text{Ind}(D)}\}$, 可辨识函数 $f(C) = \prod \sum \alpha(x, y)$, 对于 $\forall \alpha(x, y) = \{a_1, a_2, \dots, a_k\} \neq \Phi$, 令 $\sum \alpha(x, y) = a_1 \vee a_2 \vee \dots \vee a_k$;

step4: 化简可辨识函数 $f(C)$, 计算关于条件属性集 C 的 D -约简, 及最佳约简。

step5: 利用最佳约简重构 C 的 D -可辨识矩阵, 并计算相对可辨识函数 $f_i(C')$, $i = 1, 2, \dots, n = \text{card}(U)$, 其中, $f_i(C') = \prod \sum \alpha(x, y)$ 。

step6: 化简 $f_i(C')$, 计算关于条件属性 C' 的值的 D -约简, 得到分类规则;

step7: 合并相同规则, 得通用分类规则;

step8: 规则还原, 并表示。

3 实验分析

实验 1 有采自海南某地区的 32 个干酪根样本^[13], 经专家鉴定: III 型(腐植型)样本 14 个, II 2 型(腐泥-腐植型)样本 7 个, II 1 型(腐植-腐泥型)样本 11 个, 分别测试了每个样本的 9 项指标。选择 9 个不同类型的样本(每类选取 3 个)构成训练数据, 来挖掘样本类型与其相应属性(指标)的关系, 从而得到相应的分类规则。其余样本作为测试数据。训练数据集见表 1。其中 9 项测试指标作为条件属性, 依次表示为 $\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9\}$, 表中样本的实际类别为决策属性

$D = \{d\}$ 。决策属性编码为 $d = \begin{cases} 1 & \text{II 1} \\ 2 & \text{II 2} \\ 3 & \text{III} \end{cases}$

用上述基于 GA 的属性离散化算法计算每一个属性的最佳分段点, 并分别为每一区域编码, 见表 2。这样, 训练数据集量化为表 3。用上述基于 RST 的分类规则挖掘算法对表 3 进行分析, 得分类规则如下:

① IF($a_4 > 12.5$) THEN $d = \text{III 型}$;

② IF($a_3 > 32.25$) AND ($a_4 < 12.5$) THEN $d = \text{II 2 型}$;

③ IF($a_3 < 32.25$) THEN $d = \text{II 1 型}$ 。

用这些规则对测试数据进行分类, 正确率为 100%。

实验 2 表 4 的数据是卫星电源分析系统故障模拟实验台产生的数据^[14], 表中 1 代表电源损耗器故障, 2 代表电源控制器线性分流故障, 3 代表无故障。条件属性 4 个, 决策属性 1 个。随机选择其中 2/3 的数据构成训练数据集, 其余 1/3 作为测试数据。表 5 为训练数据集, 应用 2.1 节提出的基于 GA 的属性离散化算法计算每一个属性的最佳分段点, 并分别为每一区域编码, 见表 6。编码后的训练数据集见表 7。用 2.2 节基于 RST 的分类规则挖掘算法对表 7 进行分析, 得知属性是冗余属性, 与文^[14]根据专家建议分析的结果一致。剔除冗余属性从该训练数据集提取的逻辑规则如下:

① IF($Icna \leq 1.42$) AND ($Isc > 0.31$) THEN 电源损耗器

故障;

② IF($Icna > 1.46$) AND ($Isc > 0.31$) THEN 电源控制器线性分流故障;

③ IF($Isc \leq 0.31$) THEN 无故障。

④ IF($Icna > 1.42$) AND ($Icna \leq 1.46$) THEN 无故障。

使用这些分类规则对卫星电源分析系统故障模拟实验台产生的任一数据进行检验, 分类结果与实际情况完全一致, 分类正确率 100%。

表 1 训练数据集

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	d
0	25.0	10.0	40.0	25.0	15.2	8.13	2.54	9.1	35.8	III
1	5.0	30.0	50.0	15.0	14.3	7.55	0.79	4.4	5.6	III
2	10.0	20.0	40.0	30.0	19.3	10.18	1.93	6.2	32.1	III
3	45.0	10.0	40.0	5.0	10.0	1.55	38.54	21.9	5.7	II 2
4	35.0	15.0	50.0	0.0	18.0	10.79	5.68	20.3	36.9	II 2
5	30.0	25.0	45.0	0.0	9.8	2.35	37.60	12.2	3.2	II 2
6	55.0	10.0	25.0	10.0	13.1	4.95	34.08	31.4	9.2	II 1
7	65.0	10.0	25.0	0.0	11.6	2.41	39.29	29.9	7.6	II 1
8	60.0	0.0	25.0	5.0	13.1	2.09	594.80	43.7	6.8	II 1

表 2 条件属性离散化参数表

Attribute	Codes	
	1	2
a_1	[0~40.0]	(40.0~+∞)
a_2	[0~17.5]	(17.5~+∞)
a_3	[0~32.25]	(32.25~+∞)
a_4	[0~12.5]	(12.5~+∞)
a_5	[0~14.75]	(14.75~+∞)
a_6	[0~6.25]	(6.25~+∞)
a_7	[0~317.05]	(317.05~+∞)
a_8	[0~25.9]	(25.9~+∞)
a_9	[0~20.65]	(20.65~+∞)

表 3 编码后的训练数据集

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	d
0	1	1	2	2	2	2	1	1	2	1
1	1	2	2	2	1	2	1	1	1	1
2	1	2	2	2	2	2	1	1	2	1
3	2	1	2	1	1	1	1	1	1	2
4	1	1	2	1	2	2	1	1	2	2
5	1	2	2	1	1	1	1	1	1	2
6	2	1	1	1	1	1	1	2	1	3
7	2	1	1	1	1	1	1	2	1	3
8	2	1	1	1	1	1	2	2	1	3

表 4 卫星电源分析系统数据

	$Icna$	$Icna$	Vdl	Isc	$Fault(d)$
x_0	0.18	0.23	30.5	0.32	1
x_1	0.17	0.21	33.6	0.32	1
x_2	2.97	2.81	31.6	0.32	2
x_3	2.15	1.86	30.8	0.32	3
x_4	0.16	0.19	34.3	0.32	1
x_5	0.18	0.22	29.7	0.29	3
x_6	2.66	2.97	33.9	0.32	2
x_7	2.69	2.84	34.6	0.28	3
x_8	2.93	2.78	30.8	0.30	3
x_9	2.85	2.68	33.6	0.32	2
x_{10}	0.22	0.19	32.5	0.32	1
x_{11}	2.76	2.83	34.2	0.32	2
x_{12}	0.22	0.18	32.1	0.32	1
x_{13}	2.85	0.22	28.6	0.32	3
x_{14}	2.78	2.67	34.7	0.32	2

(下转第 156 页)

的颜色 R 、 G 、 B 分量, R_m 、 G_m 、 B_m 为第 i 个像素值被归并到第 m 类的颜色 R 、 G 、 B 分量。色差 $s = \sum_{i=1}^N (c_1 \Delta R_m^2 + c_2 \Delta G_m^2 + c_3 \Delta B_m^2)$, 其中 N 为图像的像素总数, c_1 、 c_2 、 c_3 为各分量的权, ΔR_m 、 ΔG_m 、 ΔB_m 为第 i 个像素值被归并到第 m 类所引起的 R 、 G 、 B 的变化值, $\Delta R_m = R_i - R_m$, $\Delta G_m = G_i - G_m$, $\Delta B_m = B_i - B_m$ 。

当 s 的值较大时, 说明图像分色后的失真度较大; 反之, 当 s 的值较小时, 说明图像分色后的失真度较小, 分色的效果好。

采用传统分色方法的图 2 失真值 s 为 7.44×10^9 , 采用改进聚类分色方法的图 4 失真值为 1.17×10^9 , 显而易见原样稿彩色图像经过改进聚类分色后图像的失真度较小, 与原图像的实际风格一致。

结束语 改进聚类分色方法与传统自动分色方法是两种不同的处理彩色像景的方法, 比较而言, 传统自动分色方法, 不管何种彩色图像, 采取同一的按频率大小处理, 方案单一, 分色后图像的颜色层次不分明。改进的 K-means 聚类算法通过采用最大频度与类内最小距离最大相结合的方法选取初始类代表点; 然后进行 K-means 聚类算法从而得到了令人满

(上接第 150 页)

表 5 训练数据集

	Icna	Icnb	Vdl	Isc	Fault(d)
x_0	0.18	0.23	30.5	0.32	1
x_1	0.17	0.21	33.6	0.32	1
x_4	0.16	0.19	34.3	0.32	1
x_6	2.66	2.97	33.9	0.32	2
x_9	2.85	2.68	33.6	0.32	2
x_{11}	2.76	2.83	34.2	0.32	2
x_3	2.15	1.86	30.8	0.32	3
x_5	0.18	0.22	29.7	0.29	3
x_8	2.93	2.78	30.8	0.30	3

表 6 条件属性量化参数表

Attribute	Codes	
	1	2
Icna	[0~1.42]	(1.42~+∞)
Icnb	[0~1.46]	(1.46~+∞)
Vdl	[0~32.2]	(32.2~+∞)
Isc	[0~0.31]	(0.31~+∞)

表 7 编码后的训练数据集

	Icna	Icnb	Vdl	Isc	Fault(d)
x_0	1	1	1	2	1
x_1	1	1	2	2	1
x_4	1	1	2	2	1
x_6	2	2	2	2	2
x_9	2	2	2	2	2
x_{11}	2	2	2	2	2
x_3	2	2	1	2	3
x_5	1	1	1	1	3
x_8	2	2	1	1	3

结束语 本文提出的基于 GA 与 RST 的分类规则挖掘算法, 利用遗传 C 均值算法实现了对样本连续型属性在没有任何领域知识可供借鉴情况下的离散化, 解决了分类规则挖

意的量化效果。失真度指标能反映分色后彩色像景图像的失真度, 通过实验表明, 改进后的聚类分析分色方法具有比传统自动分色方法较小的失真度。在彩色像景织物的开发研究中, 该分色方法在织物保持原稿的风格中起到了很好的效果。

参 考 文 献

- 李加林, 陶永政. 数码仿真彩色丝织技术及其应用[J]. 纺织学报, 2004, 25 (1): 123~125
- 施国生, 梁道雷. 黑白像景织物的图像分色方法的研究. 纺织学报[J], 2003, 24 (5): 45~46
- Jain A K, Dubes R C. Algorithm for clustering data[M]. Prentice-Hall, 1988
- Linde Y, Buzo A, Gray R M. An algorithm for vector quantizer design[J]. IEEE Trans Commun, 1980, COMM-28: 84~89
- 戚飞虎, 等译. 模式识别与图像处理[M]. 上海: 上海交通大学出版社, 1990
- 凌玲. 彩色图像量化方法的研究[J]. 华南理工大学学报, 2000, 28 (1): 81~84
- 郭欣, 赵淑清. 概率模型下的 K-means 算法在 SAR 图像分类中的应用[J]. 遥感技术与应用, 2005, 20 (2): 295~298

掘时样本数据预处理问题; 同时利用 RST 的优点, 提出了基于 RST 的分类规则挖掘方法, 该方法可挖掘出通用的分类规则。该基于 GA 与 RST 分类规则挖掘算法可用于各种数量型数据的分类规则挖掘, 比如卫星数据分类分析, 临床医学数据分析等等。但是数据预处理的速度在数据量巨大时, 还有待提高。

参 考 文 献

- 邢乃宁, 孙志挥. 一种基于粗集理论的分类规则挖掘的实现方法. 计算机应用, 2001, 21(12): 29~31
- Tsumoto S. Automated extraction of hierarchical decision rules from clinical databases using rough set model. Expert Systems with Applications, 2003, 24(2): 189~197
- Dimtras A I, Slowinski R, et al. Business failure prediction using rough sets. European Journal of Operational Research, 1999, 114 (2): 263~280
- Tay Francis E H, Shen L X. Economic and financial prediction using rough sets model, Computing, Artificial Intelligence and Information Technology. European Journal of Operational Research, 2002, 141(3): 641~659
- Peters J F, Suraj Z, Shan S, et al. Classification of meteorological volumetric radar data using rough set methods. Pattern Recognition Letters, 2003, 24(6): 911~920
- Pawlak Z. Rough set approach to knowledge-based decision support. European Journal of Operational Research, 1997, 99(1): 48~57
- Pawlak Z. Rough sets and intelligent data analysis. Information sciences, 2002, 147(1-4): 1~12
- 边肇祺. 模式识别(第二版). 北京: 清华大学出版社, 2000
- 周明, 孙树栋. 遗传算法原理及应用. 北京: 国防工业出版社, 2001
- Pawlak Z. Rough Sets. Int. J. of Computer and Information Sciences, 1982, 11(5): 341~356
- 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法. 北京: 科学出版社, 2003
- Walczak B, Massart, D L. Rough sets theory. Chemometrics and Intelligent Laboratory System, 1999, 47(1): 1~16
- 陈遵德. Rough Set 神经网络智能系统及其应用. 模式识别与人工智能, 1999, 12(3): 1~5
- 于勇, 姜兴渭, 黄文虎. 基于粗糙集理论和遗传算法的知识获取方法研究. 宇航学报, 2001, 22(3): 114~118