基于粗集理论的决策树在信用卡发放中的应用

胡来丰,舒 兰
(电子科技大学 数学科学学院,四川 成都 611731)

摘 要:基于粗集和决策树两种方法的各自优势互补,提出将粗集与决策树相结合的新方法,并将此算法运用到个人信用卡发放模型中。首先利用布尔推理算法将连续属性进行离散化处理,然后采用一种以加权和属性重要度为启发信息进行属性约简,得到降维数据,最后采用 J48 决策树算法,得到决策规则。通过对比 K 最近邻分类、朴素贝叶斯、RBF 神经网络、支持向量机等算法,这种新的数据挖掘算法保留了原有数据特点,加快了知识获取的进程,提高了模型的交叉验证率,简化了规则,取得了满意的研究结果。

关键词:粗糙集;属性约简;J48 决策树;交叉验证率
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2015)03-0142-04
doi:10.3969/j.issn.1673-629X.2015.03.032

Application of Decision Tree Based on Rough Set Theory in Credit Card Payment

HU Lai-feng, SHU Lan
(School of Mathematical Science, University of Electronic Science and Technology, Chengdu 611731, China)

Abstract: Based on the complementary advantages of rough set and decision tree method, put forward a new method of combining rough set and decision tree, and use this algorithm to personal credit card payment model. Firstly, use Boolean reasoning algorithm to continuous attribute for discretization processing, and apply a heuristic information of weighted and attribute importance for attribute reduction to obtain data with dimensionality reduction, finally utilize J48 decision tree algorithm to get the decision rules. Compared with the K nearest neighbor classification, Naive Bayes, RBF neural networks, support vector machines and other types of algorithms, this new data mining algorithm retains the original data characteristics to accelerate the process of knowledge acquisition, improving cross-model verification rate, simplifying the rules, getting the satisfactory results.

Key words: rough set; attribute reduction; J48 decision tree; cross validation rate

0 引 言

随着信用卡业务的发展,银行积累了大量的客户数据,但如何从这些数据中挖掘出信用卡的发放规则,成为一个研究热点。传统的个人信用评分[1],采取某种数量分析方法,得到每个属性的比重,并对客户的资料属性进行评分,最后加权得到总分,总分高的同意发卡,低的拒绝。但由于银行发放信用卡的标准中有一些属性是离散的,且客户申请的数据时有缺失,给银行发放信用卡带来困难,引起了研究者的关注。

粗糙集[2]是由波兰数学家 Z. Pawlak 在 1982 年提出的,能够处理不精确、不完整数据,在机器学习、数据挖掘等领域中得到了广泛的应用。粗糙集理论的核心内容之一是属性约简,就是在保持决策表分类能力不变情况下,删除其中不重要或冗余的属性,起到降维作用,从而解决了维数爆炸问题。

决策树算法[3]是一种典型的分类算法,具有分类精度高、生成的模式简单、对噪声数据有很好的健壮性等优点。1960 开始,学者相继提出了 ID3 算法[4]、C4.5 算法[5]以及在它们基础上的改进算法,使得决策树算法能够处理不同类型的数据,应用也更加广泛。

基于各自的优点,文中将两者相结合,运用到信用卡的发放决策模型中。首先运用粗糙集属性对原始数据进行约简,将约简后的数据带到决策树算法中建立

收稿日期:2014-04-26 **修回日期**:2014-07-29 **网络出版时间**:2015-01-20
基金项目:国家自然科学基金资助项目(11071178)
作者简介:胡来丰(1989-),男,硕士研究生,研究方向为不确定性的数学理论及其应用;舒 兰,教授,博士生导师,研究方向为不确定性的数学理论及其应用。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20150120.2202.032.html

决策规则,并与其他的方法进行比较说明了该方法的可用性。

1　粗糙集理论

1.1　粗糙集概念

粗糙集理论是一种新的处理模糊和不确定性知识的数学工具。其主要思想是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则。粗糙集引入了知识表达系统的概念,用一个四元组 $S=(U,A,V,f)$ 来表示。其中, $U=\{x_1,x_2,\cdots,x_n\}$ 为论域, A 是属性的非空有限集合, V 是属性集 A 的值域, $f:U\times A\rightarrow V$ 是一个信息函数。如果属性集 $A=C\cup D$, $C\cap D=\varnothing$,则 C 称为条件属性, D 称为决策属性,具有条件属性和决策属性的知识表示系统称为决策表。

定义 1:设 $S=(U,A,V,f)$ 是一个信息系统, $\forall P\subseteq A$,定义 P 上的等价关系 $\mathrm{IND}(P)=\{(x,y)\in U\times U,\forall p\in P,f(x,p)=f(y,p)\}$, $U/\mathrm{IND}(P)$ 表示与等价关系族 P 相关的知识,简记为 U/P 。

定义 2:设 $X\subseteq U$ 为论域的一个集合, $R\subseteq A$,则 X 关于 R 的下近似 $\underline{RX}=\{x\in U\mid [x]_R\subseteq X\}$, $\underline{RX}$ 是根据知识 R , U 中所有一定能归入 X 的元素的集合,称为 X 的正域,记作 $\mathrm{pos}_R(X)$ 。

定义 3:设 $S=(U,A,V,f)$ 是信息系统, $P,Q\in A$, $R\in P$, $\mathrm{pos}_P(Q)=\mathrm{pos}_{P-|R|}(Q)$,则称 R 为 P 中 Q 不必要的;否则 R 为 P 中 Q 必要的。如果 P 中的每个 R 都为 Q 必要的,则称 R 为 P 中 Q 独立的。设 $S\subseteq P$, S 是 P 的 Q 独立子集且 $\mathrm{pos}_S(Q)=\mathrm{pos}_P(Q)$,则称 S 为 P 的 Q 约简。 P 中所有 Q 必要的原始关系构成的集合成为 P 的 Q 核,记为 $\mathrm{core}_Q(P)$ 。

1.2　基于变精度加权和属性重要度粗糙集属性约简

信息系统中属性并不是同等重要的,甚至其中有些属性是冗余的,所以要对该信息系统进行属性约简。所谓属性约简,就是在保持知识库分类能力不变的条件下,删除其中不相关或不重要的属性。然而属性约简是 NP 难问题,常见的启发式属性约简算法有:基于属性重要度的属性约简、基于差别矩阵的属性约简、基于信息熵的属性约简和基于遗传算法的属性约简[6]。文中选取以属性重要度为启发信息,寻求最优或次优约简。然而当前的基于属性重要度的约简算法存在以下几个问题:属性约简的精度、属性重要度的度量方式不完备、无法区分未处理属性重要度的属性集。对于不完备问题,在文献[7]中,已给出例子证明并解释,选择不同的启发信息会产生不同的属性约简结果;对于无法区分未处理属性重要度问题,原因是在现有算法实现过程中,根据属性重要度大小逐个选择,会出现属性重要度相同的情况,当前的算法往往选择前项,有可能会带来属性的冗余。文中提出一种变精度加权和属性重要度的属性约简算法。该算法的主要思想是,将属性依赖度与信息熵进行加权求和作为新的属性重要度,以变精度为策略对相等属性重要度进行区分。

定义 4:在决策表 S 中,决策属性 D 对条件属性 C 的依赖度定义为: $\gamma_C(D)=\sum_{i=1}^{r}\frac{|RY_i|}{|U|},Y_i\in\frac{U}{D}$ 。其中, $|\bullet|$ 表示集合包含的元素个数。

定义 5:在决策表 S 中,决策属性 D 相对于条件属性 C 的条件熵定义为: $H(D\mid C)=-\sum P(X_i)\sum P(Y_j\mid X_i)\log_2 P(Y_j\mid X_i)$ 。其中, $X=\{X_1,X_2,\cdots,X_m\}$, $Y=\{Y_1,Y_2,\cdots,Y_n\}$ 分别是 C 和 D 在 U 上导出的划分[8]。

定义 6:在决策表 S 中, $\forall B\subseteq C$, $\forall a\in C-B$,定义加权和属性重要度: $\mathrm{sig}(a,B;D)=\omega(\gamma_{B\cup\{a\}}(D)-\gamma_B(D))+(1-\omega)\frac{H(D\mid B)-H(D\mid B\cup\{a\})}{\log_2|U|}$ 。其中, ω 为近似度, $\omega=|\mathrm{pos}_B(D)|/\bar{B}(D)$,称条件属性 a 对于条件属性 B 相对于决策属性 D 的重要度。

算法描述如下:

输入:一个决策表 $S=(U,A,V,f)$,其中 $A=C\cup D$, $C\cap D=\varnothing$, $\beta=1$, $\varepsilon=0.05$;

输出:决策表 S 的属性约简 R 。

该算法步骤如下:

1)由定义 4 计算出决策属性 D 关于条件属性 C 的依赖度 $\gamma_C(D)$, $R=\varnothing$, able $=\varnothing$;

2)利用区分矩阵计算出 C 相对于 D 的核 $\mathrm{core}_D(C)$, $R\Leftarrow\mathrm{core}_D(C)$;

3)计算 $\gamma_R(D)$,如果 $\gamma_R(D)=\gamma_C(D)$,输出 R ,算法结束;否则,跳下一步;

4)若 able $=\varnothing$,对于每个 $a\in C-R$,由定义 6 计算 $\mathrm{sig}(a,R;D)$,选择使 $\mathrm{sig}(a,R;D)$ 最大的属性, able $=\{a_i\}$;

5)若 $|\mathrm{able}|=1$,则 $a\Leftarrow$able, $R\Leftarrow R\cup a$,转 3);否则,令(1) $P=$able ;(2) $\beta=\beta-\varepsilon$,对每个 $b\in P$,计算 $\gamma^{\beta}_{b\cup R}(D)$,选择使 $\gamma^{\beta}_{b\cup R}(D)$ 最大的属性 b , $p\Leftarrow b$;(3)若 $|P|>1$,转(2),否则 able $=P$,转 5);

6)输出最小约简 R 。

2　J48 决策树

分类算法,就是根据文本的特征或属性,划分到已有的类别中。常用的分类算法有决策树、朴素贝叶斯、神经网络、支持向量机等。在众多的分类方法中,决策树有易于提取显式规则、计算量相对较小、可以显示重要的决策属性和较高的分类准确率等优点。决策树算法的本质是一种贪心算法,首先对数据进行处理,利用

归纳算法生成可读的规则和决策树,然后根据决策对新数据进行分析。文中将介绍 J48 决策树算法[5],J48 决策树是在 ID3 算法基础上的改进算法,继承了 ID3 算法的优点,并在 ID3 算法基础上在以下几个方面进行了改进:

(1)用信息增益率来选择属性,克服了用信息增益选择属性时偏向选择取值多的属性的不足;

(2)在树构造过程中进行剪枝;

(3)能够完成对连续属性的离散化处理;

(4)能够对不完整数据进行处理。

所以 J48 决策树算法产生的分类规则易于理解,准确率较高。具体的算法过程如下:

(1)创建节点 N

(2)如果训练集为空,在返回节点 N 标记为 Failure

(3)如果训练集中的所有记录都属于同一个类别,则以该类别标记节点 N

(4)如果候选属性为空,则返回 N 作为叶节点,标记为训练集中最普通的类

(5)for each 候选属性 attribute list

(6)if 候选属性是联系的 then

(7)对该属性进行离散化

(8)选择候选属性 attribute list 中具有最高信息增益的属性 D

(9)标记节点 N 为属性 D

(10)for each 属性 D 的一致值 d

(11)由节点 N 长出一个条件为 $D = d$ 的分支

(12)设 s 是训练集中 $D = d$ 的训练样本的集合

(13)if s 为空

(14)加上一个树叶,标记为训练集中最普通的类

(15)else 加上一个有 J48($R - \{D\}$,C,s) 返回的点[9]

3 基于粗糙集的决策树方法在信用卡发放模型的应用

基于粗糙集和决策树相结合的过程描述如下:对于给定数据,先转换为粗糙集可处理的形式(离散数据),构成一个信息决策系统。利用上文提出的属性重要度作为启发信息,不断从条件属性 C 中取出相对于决策属性 D 属性重要度的属性,与核构成新的属性集,不断重复该过程,得到最小属性约简集,将属性约简之后的数据利用 J48 决策树算法,得到一棵决策树与可读的规则。文中将此算法运用到信用卡的发放模型中,取得了满意的研究成果。

文中选取了 2004 年北京工业大学数学建模竞赛复赛试题中的数据,有样本数据 600 组。由数据表可知该决策表有 $A_1 \sim A_{15}$ 这 15 个条件属性,其中 A_2、A_3、A_8、A_{11}、A_{14}、A_{15} 为连续型变量,其余的为离散型变量。一个决策属性 d,为离散型变量,$d(-)$ 表示发放信用卡,$d(+)$ 表示不发放信用卡。

3.1 数据预处理

该数据中既包含了连续属性,又包含了离散属性。对于连续属性的处理通常采用统计分析法,但统计方法不能适用于处理离散属性。粗糙集可以处理离散属性,但不能直接处理连续属性,需要将连续属性离散化,把连续属性的取值范围或取值区间划分成数目不太多的小区间。文中用布尔推理离散方法[10]对 600 组数据里的连续属性进行离散化,得到离散化后的数据,能够被粗糙集进行处理。

3.2 属性约简

经过连续属性的离散处理及预处理之后,决策表变化结果见表 1。

表 1 离散后的数据

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	D
1	2	6	1	1	1	8	1	5	1	1	2	0	1	4	1	0
2	1	8	3	1	1	7	2	6	1	1	3	0	1	1	3	0
…									…							
599	2	8	4	1	1	9	2	7	1	1	3	0	1	1	1	0
600	2	3	1	1	1	8	1	3	0	1	2	0	1	5	3	1

运用上文提到的基于变精度加权和属性重要度的属性约简算法过程,在 Matlab 中进行编程,将离散后的数据(见表 1)带入。

利用区分矩阵求出其核为 A_2, A_3, A_4, A_8, A_9, A_{11}, 再根据文中提出的新的属性重要度约简算法,留下属性重要度大的, 删除冗余的或属性重要度小的, 最后得到最优约简为 A_2, A_3, A_4, A_7, A_8, A_9, A_{11}, A_{14}, A_{15}。

3.3 基于属性约简的 J48 决策树

将未进行粗糙集属性约简的数据带入到上面提到的 J48 决策树算法中,得到一棵叶子树为 18,大小为 26 的树,交叉验证的正确率为 86.284 7%,时间为 1.9 s,且得到的决策规则繁琐,不易解读。文中将属性约简后剩下的决策表带入 J48 决策树算法的程序中,得到一棵叶子树为 12,大小为 18 的树,交叉验证的正确率为 86.233 3%,时间为 0.6 s,得到的树见图 1。

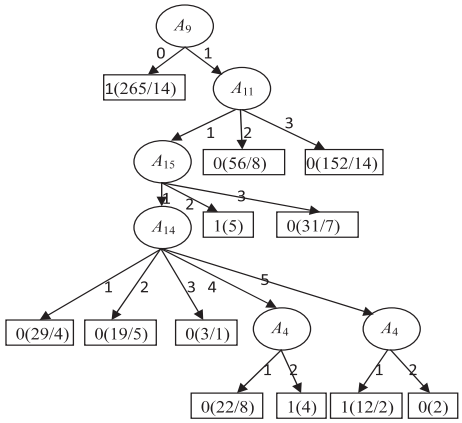

图 1　约简后的 J48 决策树

图 1 的解释:对于 A_9 属性,只要满足“0” 类别,直接可以认为能发放信用卡,其中在 265 组数据中有 14 个划分错误;属性 A_9 取“0” 类,属性 A_{15} 取在[* ,1) 区间时,属性 A_{14} 取在区间[* ,98) 时不发放信用卡,其中 29 组数据中有 4 个划分错误。

通过上面的决策树,可以很直观的理解,对于增加样本进行预测时,可以根据上面的决策树进行很好的预测。

3.4　与其他方法的比较

K 最近邻算法[11],计算测试样本到各训练样本的距离,取其中最小的 K 个,并根据这 K 个训练样本的类别得到测试样本的类别,其计算量较大,且正确率也较低;贝叶斯理论[12],能够处理不确定性信息,基于概率统计理论,在分类过程中需要知道其总体或样本的概率分布,为了获得它们,就需要大量样本;RBF 神经网络[13],具有最佳逼近性能和全局最优特性,并且结构简单,训练速度快,但没能力来解释自己的推理过程和推理依据;支持向量机方法[14]是建立在统计和结构风险最小原理基础上的,根据有限的样本信息寻求最佳分类,然而 SVM 借助二次规划来求解支持向量,求解二次规划将涉及 m 阶矩阵,m 很大时耗内存和时间。在 weka 软件上利用上述方法对信用卡数据进行分类建模,结果比较见表 2。

表 2　7 种算法结果比较

算法名称	交叉验证率/%	时间/s
K 最近邻分类	81.423 6	0.2
朴素贝叶斯分类	78.298 6	1.6
RBF 神经网络	85.416 7	2.2
支持向量机(SVM)	86.284 7	4.34
ID3 决策树	76.909 7	1.3
J48 决策树算法	86.284 7	1.9
粗糙集+J48 算法	86.233 3	0.6

从表 2 的几种算法比较来看,单从交叉验证率来说,J48 决策树分类算法与支持向量机的正确率最高,为 86.284 7%,但 J48 决策树算法在时间上节省了 20 倍。文中的基于粗糙集的 J48 决策树算法在交叉验证率上相对于 J48 决策树有略微下降,主要由于在属性约简后相对于原数据信息有所减少(用 9 个属性代替了原来的 15 个属性),但是在时间上提升了 3 倍,且在树的构造上也减少了叶子节点,减小了树的大小,得到更简单的规则,使对该决策表的分类更易直观的理解。

4　结束语

文中建立了基于粗集的决策树方法在银行信用卡发放的模型,属性约简算法采用了变精度加权和属性重要度算法,保证了属性重要度算法的完备性与精确性,得到最小约简,并与 J48 决策树相结合,得到决策规则。该方法相对于其他的分类方法,保持了较高的正确率,在时间上也大大减少,得出的决策树规则也便于直观理解。该模型也可以推广到其他模式识别问题,具有一定理论意义和推广价值。

参考文献:

[1] 姜明辉. 商业银行个人信用评估组合预测方法研究[D]. 哈尔滨:哈尔滨工业大学,2006.

[2] Pawlak Z. Rough sets[J]. International Journal of Computer and Information Sciences,1982,11(5):341-356.

[3] Quinlan J R. Induction of decision trees[J]. Machine Learning,1986,1(1):81-106.

[4] 李　霞. ID3 分类算法在银行客户流失中的应用研究[J]. 计算机技术与发展,2009,19(3):158-160.

[5] Quinlan J R. Improved use of continuous attributes in C4.5[J]. Journal of Artificial Intelligence Research,1996,4:77-90.

[6] 陶　志,许宝栋,汪定伟,等. 基于遗传算法的粗糙集知识约简方法[J]. 系统工程,2003,21(4):116-122.

[7] 洪　菁,陆金桂,石　峰. 基于改进的属性重要度的启发式算法[J]. 微计算机信息,2006,22(3-3):246-248.

[8] 苗夺谦,王　珏. 粗糙集理论中知识粗糙性与信息熵关系的讨论[J]. 模式识别与人工智能,1998,11(1):34-40.

[9] 程克非,程　蕾,黄永东. 基于 J48 决策树算法的水质评价方法[J]. 计算机工程,2012,38(11):264-267.

[10] 李　慧,闫德勤,韩　丽. 一种基于粗糙集理论的连续属性离散化新算法[J]. 计算机应用研究,2010,27(1):77-78.

[11] Zhang Bin,Srihari S N. Fast k-Nearest neighbor classification using cluster-based trees[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(4):525-528.

[12] 张亚萍,胡学钢. 基于 K-means 的朴素贝叶斯分类算法的研究[J]. 计算机技术与发展,2007,17(11):33-35.

[13] Ridella S,Rovetta S,Zunino R. Circular back propagation networks for classification[J]. IEEE Transactions on Neural Networks,1997,8(1):84-97.

[14] 刘江华,程君实,陈佳品. 支持向量机的训练算法综述[J]. 信息与控制,2002,31(1):45-50.

基于粗集理论的决策树在信用卡发放中的应用

作者： 胡来丰， 舒兰， HU Lai-feng， SHU Lan
作者单位： 电子科技大学 数学科学学院,四川 成都,611731
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(3)