Journal of System Simulation

Volume 27 | Issue 2 Article 20

9-2-2020

Attribute Reduction Algorithm on Balancing Profit and Risk

Dexing Wang
College of Information Technology, Shanghai Ocean University, Shanghai 201306, China;

Jielong Xu
College of Information Technology, Shanghai Ocean University, Shanghai 201306, China;

Hongchun Yuan
College of Information Technology, Shanghai Ocean University, Shanghai 201306, China;

Attribute Reduction Algorithm on Balancing Profit and Risk

Abstract

Abstract: Usually it is taken for granted to achieve the maximal profit at the cost of the minimal risk. It is an important problem of how to balance profit and risk, considering introducing profit and risk to attribute reduction so as to find practical algorithms in decision-making process. *A decision-theoretic model was built, which could balance profit and risk combining with decision-theoretic rough set model and minimum risk of Bayes decision and find optimal combinations of risk in a certain level of expected profit, then a heuristic algorithm of attribute reduction was proposed*, which took the function of balancing profit and risk as the target of heuristic attribute reduction, and it could reduce the scales and complexity of data model, and then improve the simulation precision of the model system.

Keywords

attribute reduction, decision table, rough set, profit, risk

第 27 卷第 2 期 系统仿真学报© Vol. 27 No. 2
2015 年 2 月 Journal of System Simulation Feb., 2015

基于收益和风险优化的属性约简算法

王德兴, 徐节龙, 袁红春
（上海海洋大学信息学院, 上海 201306）

摘要：决策者总是希望实现收益最大而承担的风险最小，如何平衡或兼顾两者，可考虑引入收益和风险因素进行属性约简以便做出寻找有效的、切实可行的决策。*在一定的预期收益水平下通过优化组合收益和风险，结合粗糙集和贝叶斯模型，建立了收益和风险优化的决策模型*，以每个属性的收益风险平衡组合函数作为指标进行启发式属性约简，该算法减少数据模型的规模和复杂度，并提高模型系统的仿真精度。

关键词：属性约简；决策表；粗糙集；收益；风险
中图分类号：TP301.6 文献标识码：A 文章编号：1004-731X (2015) 02-0369-07

Attribute Reduction Algorithm on Balancing Profit and Risk

Wang Dexing, Xu Jielong, Yuan Hongchun
(College of Information Technology, Shanghai Ocean University, Shanghai 201306, China)

Abstract: Usually it is taken for granted to achieve the maximal profit at the cost of the minimal risk. It is an important problem of how to balance profit and risk, considering introducing profit and risk to attribute reduction so as to find practical algorithms in decision-making process. *A decision-theoretic model was built, which could balance profit and risk combining with decision-theoretic rough set model and minimum risk of Bayes decision and find optimal combinations of risk in a certain level of expected profit, then a heuristic algorithm of attribute reduction was proposed*, which took the function of balancing profit and risk as the target of heuristic attribute reduction, and it could reduce the scales and complexity of data model, and then improve the simulation precision of the model system.

Keywords: attribute reduction; decision table; rough set; profit; risk

引言

粗糙集理论是 Z Pawlak 教授提出的一种处理模糊和不确定性知识的数学工具[1]，已被广泛应用到知识获取、决策分析、模式识别和数据挖掘等领域[2-4]。属性约简是粗糙集理论中一个重要的研究方向，属性约简是在保持知识库的分类或决策能力不变的条件下，删除其中不相关或者不重要的知识。常用的启发式算法有基于可分辨矩阵的属性约简算法[5-7]、基于变精度的属性约简算法[8-9]、基于概念的属性约简算法[10]和基于决策风险的属性约简算法[11-16]等。然而，Wong S K M 和 Ziarko W 已经证明最小属性的约简是 NP-hard 问题[17]。

文献[5]提出一种带启发式知识的差别矩阵的属性约简算法，该算法能在很大程度上找到决策表的最小属性约简；文献[6]提出基于样例选取的差别矩阵属性约简算法，该算法能有效地减少大型数据库的存储空间和算法的执行时间；文献[7]提出一种基于简化差别矩阵的增量式属性约简算法；文

作者简介：王德兴(1968-)，男，河北，博士，副教授，研究方向为人工智能、数据挖掘；徐节龙(1988-)，男，安徽，硕士，研究方向为粗糙集、概念格；袁红春(1971-)，男，江苏，博士，教授，研究方向为人工神经网络、智能计算。
收稿日期：2013-08-15 修回日期：2014-09-23
基金项目：上海市科委科研计划重点支撑资助项目(12510502000)；上海市科委科技支撑项目(14391901400)

献[8]提出将变精度粗糙集的属性约简用于集成神经网络中，以提高其分类性能；文献[9]针对变精度粗糙集的属性约简问题提出基于容差矩阵和属性核求解变精度粗糙集最小属性约简算法。文献[10]借助内涵亏值的思想证明在每个唯一上近邻概念的内涵亏值中各任取一个元素组成的集合就是一个约简，给出了一个求全部约简的算法；文献[11]在决策粗糙集模型下首先定义了保持决策正域不变的属性约简，并提出了一种泛化的属性约简定义，约简是保持或提高在原有条件属性集下某种度量值的最小属性集合；在文献[11]的基础上，Yao 等从三枝决策语义角度定义了保持决策不变的属性约简和决策发生变化时的属性约简，并论证了概率三枝决策优于概率二枝决策和定性的三枝决策[12]；文献[13]探讨决策粗糙集风险函数的语义学，并用不同的风险替换实际的风险，构造一种新的概率规则选择标准；文献[14-16]提出基于风险最小化的决策粗糙集模型的属性约简，并给出最优化问题的属性约简算法。

通过属性约简可以在海量的、复杂数据中挖掘其中隐藏的属性间关联，从而简化数据模型并提高其系统模型的仿真精度。同时减少数据模型的规模和复杂度，也便于进行模型系统的计算机仿真。一般情况下，将决策问题划分为 2 种类型：确定型、不确定型，而且针对每一类型的决策问题，都有比较成熟的理论和算法，从而形成整个决策体系。风险型决策问题是不确定性决策论的核心组成部分，期望收益最大和风险最小准则是用户追求的目标。但是这种理想状况由于实际问题比较复杂或存在噪音的干扰无法实现，因此如何兼顾两者，可考虑引入收益和风险因素进行属性约简以便做出寻找有效的、切实可行的决策。针对上述问题，根据尽可能的风险最小、收益最大的思想，结合决策粗糙集模型和收益、风险优化的贝叶斯决策，提出了一种基于收益和风险优化的属性约简算法，探索一种能得到收益和风险相对较优的决策。

本文组织结构如下：第 1 部分介绍决策粗糙集模型的相关知识；第 2 部分探讨决策粗糙集的属性约简及其算法；第 3 部分是应用实例分析，并与常见典型属性约简算法进行比较；第 4 部分是总结。

1 相关的概念和定义

1.1 三枝决策规则

定义 1 一个决策表6可以表示为$S=(U,A,V,f)$，其中 U 为对象的非空有限集合，称之为论域；A 为属性的非空有限集合，又可分为条件属性集 C 和决策属性集 D，即 $A=C\cup D$ 且 $C\cap D\neq\varnothing$；$V=\cup_{a\in C\cup D}V_a$，V_a 是属性 a 的值域，$f:U\times A\to V$ 是信息函数，它用来指定论域 U 中每个对象各属性的取值[2]。

定义 2 对决策表 $S=(U,A,V,f)$，令 $R\subseteq A$，$ind(R)=\{(x_i,x_j)\mid f(x_i,b)=f(x_j,b),\forall b\in R\}$ 称为 S 的不可区分关系。显然不可区分关系是一个等价类，含 x 的等价类记为 $[x]_R$，R 在 U 上导出的划分记为 U/R [2]。

定义 3 设 $P=(U,R)$ 是 Pawlak 近似空间，R 是 U 上的等价关系，对于任意 $X\subseteq U$，则 X 关于近似空间 (U,R) 的下近似集和上近似集分别表示为：

$$R\_(X)=\{x\in U\mid[x]_R\subseteq X\}=\cup\{[x]_R\mid[x]_R\subseteq X\} \tag{1}$$

$$R^{-}(X)=\{x\in U\mid[x]_R\cap X\neq\varnothing\}=\cup\{[x]_R\mid[x]_R\cap X\neq\varnothing\} \tag{2}$$

三枝决策规则[12]是由 3 个不相交的区域即正区域、边界域和负区域产生。其中，正区域 $POS_R(X)=R\_(X)$，边界域 $BND_R(X)=R^{-}(X)-R\_(X)$，负区域 $NEG_R(X)=U-R^{-}(X)$。其中 $\forall x\in POS_R(X)$ 属于 X 或 $\forall x\in NEG_R(X)$ 不属于 X 都是确定的，而对于 $\forall x\in BND_R(X)$ 是否属于 X 是不确定的，因此可以得到正区域、边界域和负区域的决策规则：

$$Des([x])\to_P Des(C),[x]\subseteq POS_R(X);$$

$$Des([x])\to_B Des(C),[x]\subseteq BND_R(X);$$

$$Des([x])\to_N Des(C),[x]\subseteq NEG_R(X);$$

三枝决策规则给粗糙集理论提供了一种新的解释：正规则表示接受，负规则表示拒绝，边界规

则需有较多的知识进行探讨。

1.2 风险最小化的贝叶斯决策

在决策论中，决策对象所有可能的状态 $\omega_j(j=1,2,\ldots,\mathrm{s})$ 构成的状态空间 $\Omega=\{\omega_1,\omega_2,..,\omega_s\}$，决策空间表示所有可能的各种决策组合组成的集合，其中 $a_i(i=1,2,\ldots,m)$ 表示各种决策。每一个决策行为 a_i 都会带来一定的损失，其大小取决于选择的决策和决策对象的状态，因此损失可以视为决策和状态的函数，这种函数关系可以通过损失矩阵进行表述。在研究 s 个不同状态的对象和 m 个不同决策时，可以用 m×s 的表格来描述决策损失矩阵，其中损失函数 $\lambda(a_i\,|\,\omega_j)$ 表示在状态 ω_j 下，采用决策 a_i 时所可能带来的损失。

假定各个状态发生的先验概率 $P(\omega_j)$ 和状态 ω_j 下具有的描述为 $[x]$ 的条件概率 $P(x|\omega_j)$ 均已知，根据贝叶斯概率公式，在给定描述 $[x]$ 下处于状态 ω_j 的后验概率 $P(\omega_j\,|\,x)$ 为

$$P(\omega_j\,|\,x)=\frac{P(x\,|\,\omega_j)*P(\omega_j)}{P(x)} \tag{3}$$

其中，$P(x)=\sum_{j=1}^{s}P(x\,|\,\omega_j)*P(\omega_j)$，则有

$$P(\omega_j\,|\,x)=\frac{P(x\,|\,\omega_j)*P(\omega_j)}{\sum_{j=1}^{s}P(x\,|\,\omega_j)*P(\omega_j)} \tag{4}$$

由于需考虑决策所带来的总体损失，因此不能仅依据后验概率的大小进行决策，而应考虑决策是否使总体的损失最小，并以此作为选择最优决策的依据。对于给定的描述[x]，采用决策 a_i 所对应的损失值为

$$R(a_i\,|\,x)=\sum_{j=1}^{s}\lambda(a_i\,|\,\omega_j)*P(\omega_j\,|\,x) \tag{5}$$

在考虑误分类带来的风险时，我们希望最优决策能够具有最小的风险。如果在做出每个决策之前，都使得其条件风险值最小，则对所有的 x 进行决策时，其期望风险也必然最低，这就是风险最小化的贝叶斯决策。因此，风险最小化的贝叶斯决策可以形式化为

$$\text{If}\ \ R(a_k\,|\,x)=\min_{i\in\{1,2,\ldots,m\}}R(a_i\,|\,x),$$
$$\text{then}\ \ a=a_k \tag{6}$$

根据风险最小化的贝叶斯规则的定义，求解风险最小化的贝叶斯决策可以按照以下步骤来实现：

给定具体描述 $[x]$，根据式(4)计算后验概率；

根据后验概率和损失矩阵计算各个决策的期望风险，如式(5)所示；

比较各个决策的期望风险，由式(6)找出期望风险最小的决策，即风险最小化的贝叶斯决策。

1.3 决策粗糙集模型

设 $\Omega=\{\omega_1,\omega_2,\ldots,\omega_s\}$ 表示 S 个特征状态的有限集合，每个 ω_i 都是 U 的子集；$A=\{a_1,a_2,\ldots,a_m\}$ 表示 m 个可能决策行为构成的有限集合；$\lambda(a_i\,|\,\omega_j)$ 表示在状态 ω_j 下做出决策 a_i 的风险；$P(\omega_j\,|\,x)$ 表示在给定描述 $[x]$ 下处于状态 ω_j 的条件概率。因此，对象在给定的描述 $[x]$ 下采用决策 a_i 的期望风险：

$$R(a_i\,|\,[x])=\sum_{j=1}^{s}\lambda(a_i\,|\,\omega_j)*P(\omega_j\,|\,[x]) \tag{7}$$

对于给定的描述 $[x]$，记 $\tau(x)$ 为一个决策规则，即 $\tau(x)\in A$，则 τ 是描述空间到 A 的一个函数。此时，决策规则的风险为：

$$R=\sum_{[x]}R(\tau(x)\,|\,[x])*P([x]) \tag{8}$$

其中式(8)是对整个决策系统而言的。显然，如果决策规则 $\tau(x)$ 使得对于每个 $[x]$ 而言条件风险 $R(\tau(x)\,|\,[x])$ 尽可能的小，那么总体风险就能达到最小值[12]。

在决策粗糙集模型中，为简化问题考虑具有互补的 2 种状态的状态集合 $\Omega=\{\omega,\sim\omega\}$。可以分 2 种情形进行讨论：

情形 1 全体论域分为 2 个部分：$POS(\omega)$ 和 $NEG(\omega)$。

情形 2 全体论域分为 3 个部分：$POS(\omega)$，$BND(\omega)$ 和 $NEG(\omega)$。

对于情形 1，简单从略。对于情形 2，给定决策集 $\mathrm{A}=\{a_P,a_B,a_N\}$，其中 a_P，a_B，a_N 分别表示将对象分类为正域 $POS(\omega)$、边界域 $BND(\omega)$ 和负域 $NEG(\omega)$。λ_{PP}，λ_{BP}，λ_{NP} 分别表示为对象 x 属于概念 ω 时，做出 a_P，a_B，a_N 3 种分类决策

所对应的风险函数。λ_{PN}， λ_{BN}， λ_{NN} 分别表示为对象 x 不属于概念 ω 时做出 a_P， a_B， a_N 3 种分类决策所对应的风险函数。依次计算 3 种决策的期望风险分别为：

$R_P = R(a_P \mid [x]) = \lambda_{PP} * P(\omega \mid [x]) + \lambda_{PN} * P(\sim\omega \mid [x])$，#

$R_B = R(a_B \mid [x]) = \lambda_{BP} * P(\omega \mid [x]) + \lambda_{BN} * P(\sim\omega \mid [x])$，

$R_N = R(a_N \mid [x]) = \lambda_{NP} * P(\omega \mid [x]) + \lambda_{NN} * P(\sim\omega \mid [x])$。

根据贝叶斯最小风险决策原则，得到如下形式的决策规则：

$(P) If \quad R_P \leqslant R_B \quad and \quad R_P \leqslant R_N$，

$decide \quad x \in POS(\omega)$；

$(B) If \quad R_B \leqslant R_P \quad and \quad R_B \leqslant R_N$，

$decide \quad x \in BND(\omega)$；

$(N) If \quad R_N \leqslant R_P \quad and \quad R_N \leqslant R_B$，

$decide \quad x \in NEG(\omega)$。

现在考虑风险函数，即要求 $\lambda_{PP} \leqslant \lambda_{BP} < \lambda_{NP}$ 和 $\lambda_{NN} \leqslant \lambda_{BN} < \lambda_{PN}$，也就是当 x 属于概念 ω 时决策为正域 $POS(\omega)$ 的代价值不大于决策为边界域 $BND(\omega)$ 的代价值，两者的风险均严格地小于决策为负域 $NEG(\omega)$ 的风险；当 x 不属于概念 ω 时决策为负域 $NEG(\omega)$ 的代价值不大于决策为边界域 $BND(\omega)$ 的代价值，两者的风险均严格地小于决策为正域 $POS(\omega)$ 的风险。

根据文献[14]的推导过程，可以得到简化的决策规则：

$(P_1) If \quad P(\omega \mid [x]) \geqslant \alpha, decide \quad x \in POS(\omega)$；

$(B_1) If \quad P(\omega \mid [x]) < \alpha, decide \quad x \in BND(\omega)$；

$(N_1) If \quad P(\omega \mid [x]) \leqslant \beta, decide \quad x \in NEG(\omega)$。

根据以上决策规则，论域 U 可以通过阈值 α，β 划分为 3 个区域，即决策正域、决策边界域和决策负域，从语义上看，这 3 个区域可以对应 3 种规则类型，即正域决策规则，边界域决策规则和负域决策规则。

同理，可以定义决策规则的期望收益为：

$$E(a_i \mid [x]) = \sum_{j=1}^{s} \delta(a_i \mid \omega_j) * P(\omega_j \mid [x]) \tag{9}$$

$$E = \sum_{[x]} E(\tau(x) \mid [x]) * P([x]) \tag{10}$$

其中：$\delta(a_i \mid \omega_j)$ 为状态 ω_j 下做出决策 a_i 的收益时。

2 基于收益和风险优化的属性约简

2.1 基于收益和风险优化的决策模型

在决策表 $S = (U, A = C \cup D, V, f)$ 中，通过设定的阈值 (α, β)，可以将论域 U 划分为以下 3 个区域：

$$\begin{aligned} &\mathrm{POS}_{(\alpha,\beta)}(\pi_D \mid \pi_R) = \\ &\quad \{x \in U \mid P(D_{\max}([x]_R) \mid [x]_R) \geqslant \alpha\} \\ &BND_{(\alpha,\beta)}(\pi_D \mid \pi_R) = \\ &\quad \{x \in U \mid \beta < P(D_{\max}([x]_R) \mid [x]_R) < \alpha\} \\ &NEG_{(\alpha,\beta)}(\pi_D \mid \pi_R) = \\ &\quad \{x \in U \mid \mathrm{P}(D_{\max}([x]_R) \mid [x]_R) \leqslant \beta\}。 \end{aligned} \tag{11}$$

其中，$D_{\max}([x]_R) = \arg\max_{D_{i \in \pi D}} \{\dfrac{[x]_R \cap \pi_D}{[x]_R}\}$，表示使 D 取得最大值时的划分 $[x]$。

在收益和风险优化的决策中，需要考虑决策带来的收益和风险。令 $P = P(D_{\max}([x]_R))$，根据风险最小化的贝叶斯决策，每个决策规则的风险为：

正区域规则的风险：$(1 - P_i) * \lambda_{PN}$

边界区域规则的风险：

$$(P_i * \lambda_{BP} + (1 - P_i) * \lambda_{BN}) \tag{12}$$

负区域规则的风险：$\mathrm{P}_K * \lambda_{NP}$

同理，每个决策规则的收益为：

正区域规则的收益：$(1 - P_i) * \delta_{PN}$

边界区域规则的收益：

$$(P_i * \delta_{BP} + (1 - P_i) * \delta_{BN}) \tag{13}$$

负区域规则的收益：$P_K * \delta_{NP}$

结合以上对正区域、边界区域和负区域的期望收益和风险的推导，可归纳出决策组合的目标有：

(1) 在一定的期望收益水平下，通过决策的优化组合，将使所承担的决策风险最小；

(2) 在一定的风险水平下，经过决策的优化组合，使得其获得最大的预期收益；

(3) 寻求收益概率和风险概率的一种平衡点，承受的风险很大，获得的收益也很大。

在本文中，以目标(1)为基础建立了收益和风险优化组合的决策模型。

模型(1) 期望收益水平一定，决策风险极小化

目标函数：$\varphi(x) = \min R$

$$\text{约束条件：}\begin{cases} E \geqslant k \\ \sum P'_{x_i \in [x]} = 1 \\ P'([x]) \geqslant 0 \end{cases}$$

其中：K 表示期望收益，即约简前的收益取值。

该模型核心是在一定收益下通过分析比较各种决策的风险损失，找出收益风险优化的决策，作为把对象划分到正域、负域和边界区域的依据，由此得到区分正域、负域和边界的阈值，并根据该阈值最终形成正域决策、负域决策和边界决策。

同理，以目标(2)、目标(3)为基础构建决策模型，在此从略。由此，根据模型(1)风险和收益的定义如下：

定义 6 在决策表 $S=(U,A=C\cup D,V,f)$ 中，基于属性集 $R\subseteq C$ 的风险定义为：

$$COST_R=\sum_{P_i\in POS_{(\alpha,\beta)}(\pi_D|\pi_R)}(1-P_i)*\lambda_{PN}+\sum_{P_j\in BND_{(\alpha,\beta)}(\pi_D|\pi_R)}(P_j*\lambda_{BP}+(1-\mathrm{P}_j)*\lambda_{BN})+\sum_{P_K\in NED_{(\alpha,\beta)}(\pi_D|\pi_R)}P_K*\lambda_{NP} \quad (14)$$

定义 7 在决策表 $S=(U,A=C\cup D,V,f)$，基于属性集 $R\subseteq C$ 的期望收益定义为：

$$GAIN_R=\sum_{P_i\in POS_{(\alpha,\beta)}(\pi_D|\pi_R)}(1-P_i)*\delta_{PN}+\sum_{P_j\in BND_{(\alpha,\beta)}(\pi_D|\pi_R)}(P_j*\delta_{BP}+(1-P_j)*\delta_{BN})+\sum_{P_K\in NED_{(\alpha,\beta)}(\pi_D|\pi_R)}P_K*\delta_{NP} \quad (15)$$

定义 8 在决策表 $S=(U,A=C\cup D,V,f)$ 中，$\#R\subseteq C$ 是基于收益风险优化的属性约简的充要条件是：

(1) $R=\arg\min_{R\subseteq C}\{COST_R\},GAIN_R\geqslant GAIN_C$；

(2) 对 $\forall R'\subseteq R,COST_{R'}>COST_R$。

在定义 8 中，(1)式是其充分条件，保证在一定的期望收益下约简引入的风险最小化，其中 arg 表示 R 使 $COST_R$ 最大值时的参数 R；(2)式是其必要条件，确保得到的约简是最小约简。

定义 9 在决策表 $S=(U,A=C\cup D,V,f)$ 中，$R\subseteq C$，$b\in C\setminus R$ 在 R 中的收益风险平衡组合函数定义为：

$$b=\arg\min COST_{R\cup\{b\}},GAIN_{R\cup\{b\}}\geqslant GAIN_R \quad (16)$$

arg 表示最小值的参数，即 b 是一个能使得 $COST_{R\cup\{b\}}$ 这个函数取最小值的一个值。

2.2 基于收益和风险优化的属性约简

在模型(1)中，每个对象所处的区域是有风险最小化原则决定的，划分到哪一个区域所带来的风险最小，就将对象划分到该区域。区域的变化实质上是有风险的变化所决定的，每次决策都已经在期望收益水平一定下决策风险最小化的原则下进行，基于收益和风险优化的属性约简定义：

定义 10 在决策表 $S=(U,A=C\cup D,V,f)$ 中，条件属性集 $R\subseteq C$ 是 C 相对于 D 的一个约简的充要条件是：

(1) $POS_{(\alpha,\beta)}(\pi_D\mid\pi_R)=POS_{(\alpha,\beta)}(\pi_D\mid\pi_C)$

(2) 对 $\forall a\in R,POS_{(\alpha,\beta)}(\pi_D\mid\pi_{R-\{a\}})\neq POS_{(\alpha,\beta)}(\pi_D\mid\pi_R)$

由定义 10 可知，该约简是使得在一定的期望收益下风险最小化的最小决策属性集合，决策者基于约简后的属性集合所作的决策风险要小于等于基于约简前的属性集合所作的决策风险。

基于收益和风险优化的属性约简算法思想是比较约简前的风险和收益与约简后的风险和收益的大小，确保约简后的风险不大于约简前的风险，而约简后的收益不小于约简前的收益，在此启发式下进行属性约简。

相应的属性约简算法描述如下：

算法 1. 基于收益和风险优化的属性约简算法步骤：

输入：决策表 $S=(U,A,V,f),A=C\cup D$；

输出：决策表 S 的属性约简集 R。

Step 1：计算初始属性集 $R=\phi$；令待选子集 $G=C$;

Step 2：计算属性集 R 的期望收益 $GAIN_R$ 和属性集 C 的风险代价 $COST_C$

Step 3：while $G\neq\phi$

(1)：对每个属性 $b_i\in G$，计算其风险代价

$COST_{R\cup\{b_i\}}$ 和期望收益 $GAIN_{R\cup\{b_i\}}$；

(2)：选出收益风险平衡组合函数较优的属性：$b_i^* = \arg\min COST_{R\cup\{b_i\}}, GAIN_{R\cup\{b_i\}} \geqslant GAIN_R$；

(3)：更新 $R = R\cup\{b_i^*\}$， $\mathrm{G} = \mathrm{G} - \{b_i^*\}$；

(4)：若 $COST_R \leqslant COST_C$，则转至 Step 4；

Step 4：输出最优属性约简集 R，算法结束。

3 实例分析

为了验证本文所提出的算法的有效性，以一个如表 1 所示的决策表为例进行说明。根据文献[2]提出的经典粗糙集模型的约简算法，可以得到该决策表的正区域 $POS(\pi_D \mid \pi_R) = \{u_1, u_3, u_4, u_7\}$，其属性约简结果有：$\{c_1, c_2, c_5\}$，$\{c_2, c_3, c_5\}$，$\{c_2, c_4, c_5\}$，$\{c_2, c_5, c_6\}$，其中 $\{c_2, c_5\}$ 是属性核。根据文献[11]提出的基于属性核的最小属性约简算法可以求得表 1 的最小约简有：$\{c_2, c_4, c_5\}$，$\{c_2, c_5\}$，$\{c_4, c_5, c_6\}$，其中属性核为 $\{c_5\}$。根据文献[18]提出的基于风险最小化的属性约简算法，可以得到以下的约简结果：$\{c_1, c_2, c_5\}$，$\{c_2, c_3, c_5\}$，$\{c_2, c_4, c_5\}$，$\{c_2, c_4, c_6\}$，$\{c_2, c_5, c_6\}$，$\{c_4, c_5\}$，但是该属性约简不存在属性核。

对于表 1，设定其风险和收益的代价函数值：$\lambda_{PP} = 0$， $\lambda_{BP} = 1$， $\lambda_{NP} = 3$， $\lambda_{PN} = 6$， $\lambda_{BN} = 3$，$\lambda_{NN} = 0$， $\delta_{PP} = 4$， $\delta_{BP} = 1$， $\delta_{NP} = 0$， $\delta_{PN} = 0$，$\delta_{BN} = 3$， $\delta_{NN} = 8$，则由此我们可以计算得到 $\alpha = 0.75$， $\beta = 0.6$。

需要指出的是，每个风险和损失函数需要依赖于具体的研究领域并由该领域的专家凭借自己的经验或通过对历史数据分析从而得出具体的取值。由基于收益和风险优化的约简算法，可以计算得到所有的收益一定时，风险最小的约简结果为：$\{c_1, c_2, c_5\}$，$\{c_2, c_3, c_5\}$，$\{c_2, c_4, c_5\}$，$\{c_2, c_4, c_6\}$，$\{c_2, c_5, c_6\}$，$\{c_2, c_5\}$， $\{c_2\}$，其中属性核为 $\{c_2\}$。可以求得 $COST_{\{c_2,c_5\}} = 13$， $COST_{\{c_2\}} = 12$， $COST_{\{c_5\}} = 15$，$GAIN_{\{c_2,c_5\}} = 18$， $GAIN_{\{c_2\}} = 21$， $GAIN_{\{c_5\}} = 16$。

本算法和常见典型约简算法的比较结果见表 2。从分类质量考虑，文献[2,9]和本算法都具有属性核，而文献[16]提出的属性约简不具有属性核。

从属性约简结果的角度来看，文献[16]和本算法能得到较为充分的属性约简结果，便于挖掘出满足用户感兴趣的规则。如最小支持度阈值设为 2/9，根据本约简算法和文献[2,9]得到的属性核进行相应的规则挖掘，结果如表 3 所示。

表 1 决策表

U	C						D
	c_1	c_2	c_3	c_4	c_5	c_6	
u_1	1	1	1	1	1	1	d_1
u_2	1	1	1	0	0	1	d_1
u_3	0	0	1	1	1	0	d_2
u_4	1	0	1	0	1	1	d_2
u_5	0	0	1	1	0	1	d_2
u_6	1	1	1	0	0	1	d_3
u_7	0	1	0	1	0	0	d_3
u_8	1	1	1	0	0	1	d_3
u_9	0	0	1	1	0	1	d_3

表 2 本算法和常见典型约简算法的比较结果

约简算法	约简结果	属性核
经典的粗糙集约简算法[2]	{c_1, c_2, c_5}, {c_2, c_3, c_5}, {c_2, c_4, c_5}, {c_2, c_5, c_6}	{c_2, c_5}
基于属性核的最小属性约简算法[9]	{c_2, c_4, c_5}, {c_4, c_5, c_6}, {c_2, c_5},	{c_5}
风险最小化的属性约简算法[16]	{c_1, c_2, c_5}, {c_2, c_3, c_5}, {c_2, c_4, c_5}, {c_2, c_4, c_6}, {c_2, c_5, c_6}, {c_4, c_6}	不存在
收益和风险优化的属性约简算法	{c_1, c_2, c_5}, {c_2, c_3, c_5}, {c_2, c_4, c_5}, {c_2, c_4, c_6}, {c_2, c_5, c_6}, {c_2, c_5}	{c_2}

表 3 由属性约简核挖掘的规则

约简算法	规则	支持集	置信度
经典的粗糙集约简算法	c_2=0 and c_5=1→ d=d_2	{ u_3, u_4}	0.6667
	c_2=1 and c_5=0→ d=d_3	{ u_6, u_7, u_8}	0.75
基于属性核的最小属性约简算法	c_5=1 → d=d_2	{ u_3, u_4 }	0.6667
	c_5=0 → d=d_3	{ u_6, u_7, u_8, u_9 }	0.8
收益和风险优化的属性约简算法	c_2=1 → d=d_1 or d=d_3	{ u_1, u_2, u_6, u_7, u_8}	0.8333
	c_2=0 → d=d_2	{ u_3, u_4, u_5}	1

由以上的论述可知，对于挖掘出满足用户感兴趣的规则来说，基于收益和风险优化的属性约简具有相对较好的属性核和约简结果，便于挖掘相对较多满足用户需求的规则，因而具有较好的优势。

4 结论

基于收益和风险优化的属性约简算法是在一定的预期收益水平下通过优化组合收益和风险，结合粗糙集和贝叶斯模型，建立了收益和风险优化的决策模型, 通过本算法的属性约简可以简化数据模型，提高其系统模型的仿真精度和效率，便于进行模型系统的计算机仿真等方面具有很重要的实际应用价值。

参考文献:

[1] Pawlak Z. Rough Sets [J]. International Journal of Computing and Information Sciences (S1708-0479/1708-0460), 1982, 11(5): 341-356.

[2] 张文修, 吴伟业, 梁吉业, 等. 粗糙集理论与方法 [M]. 北京: 科学出版社, 2001.

[3] Wei Ling. Li Hongru. Zhang Wenxiu. Knowledge Reduction Based on the Equivalence Relations Defined on Attribute Set and Its Power Set [J]. Information Sciences (S0020-0255), 2007, 177(15): 3178-3185.

[4] Liang Hongli, Wang Jue, Yao Yiyu. User-oriented Feature Selection for Machine Learning [J]. The Computer Journal (S1460-2067/0010-4620), 2007, 50(4): 421-434.

[5] 蒋瑜, 王燮, 叶振. 基于差别矩阵的 Rough 集属性约简算法 [J]. 系统仿真学报, 2008, 20(14): 3717-3720, 3725. (Jiang Yu, Wang Xie, Ye Zhen. Attribute Reduction Algorithm of Rough Sets Based on Discernibility Matrix [J]. Journal of System Simulation (S1004-731X), 2008, 20(14): 3717-3720, 3725.)

[6] 王熙照, 王婷婷, 翟俊海. 基于样例选取的属性约简算法 [J]. 计算机研究与发展, 2012, 49(11): 2305-2310.

[7] 葛浩, 李龙澍, 杨传健. 基于简化差别矩阵的增量式属性约简 [J]. 四川大学学报(工程科学版), 2013, 45(1): 116-124.

[8] 周涛, 张艳宁, 陆惠玲, 等. 基于属性约简-聚类分析的集成神经网络[J]. 系统仿真学报, 2010, 22(6): 1365-1369. (Zhou Tao, Zhang, Yanning, Lu Huiling, *et al*. Ensemble Neural Network Based on Attribute Reduction and Cluster Analysis [J]. Journal of System Simulation (S1004-731X), 2010, 22(6): 1365-1369.)

[9] 陈昊, 杨俊安, 庄镇泉. 变精度粗糙集的属性核和最小属性约简算法 [J]. 计算机学报, 2012, 35(5): 1011-1017.

[10] 马垣. 形式概念中的内涵亏值及属性约简 [J]. 模式识别与人工智能, 2013, 26(12): 1096-1105.

[11] Yao Yiyu, Zhao Yan. Attribution Reduction in Decision-theoretic Rough Set Models [J]. Information Sciences (S0020-0255), 2008, 178(17): 3356-3373.

[12] Yao Yiyu. The Superiority of Three-way Decisions in Probabilistic Rough Set Models [J]. Information Sciences (S0020-0255), 2011, 181(6): 1080-1096.

[13] Liu Dun, Li Tianrui, Ruan Da. Probabilistic Model Criteria with Decision-theoretic Rough Sets [J]. Information Sciences (S0020-0255), 2011, 181(17): 3709-3722.

[14] Li Huaxiong, Zhou Xianzhong. Risk Decision Making Based on Decision-theoretic Rough Set: A Multi-view Decision Model [J]. International Journal of Computational Intelligence Systems (S1875-6883), 2011, 4(1): 1-11.

[15] 贾修一, 商琳, 陈家骏. 决策风险最小化属性约简 [J]. 计算机科学与探索, 2011, 5(2): 155-160.

[16] Jia Xiuyi, Liao Wenhe, Tang Zhenmin, *et al*. Minimum Cost Attribute Reduction in Decision-theoretic Rough Set Models [J]. Information Sciences (S0020-0255), 2013, 219(10): 151-167.

[17] Wong S K M, Ziarko W. On Optimal Decision Rules in Decision Table [J]. Bulletin of Polish Academy of Sciences, 1985, 33(11): 693-696.