基于三支决策理论的条件属性权重构造方法

薛占熬 朱泰隆 薛天宇 刘 杰
(河南师范大学计算机与信息工程学院 新乡 453007)
(智慧商务与物联网技术河南省工程实验室 新乡 453007)

摘 要 针对传统决策过程中权重规则确定的主观性和参数数值计算的不确定性问题,在粗糙集和三支决策理论的基础上,对条件属性权重构造方法进行了研究。重新定义了属性确定度和属性约简度,提出了一种属性权重构造方法,通过实例将该方法与其它条件属性权重构造方法进行了分析比较,证明了其有效性。该方法基于数据本身,不需要先验信息,从客观的角度对属性进行判断,决策者通过该方法可以得到更加合理的权重分配,做出符合实际的决策。该论文对研究属性权重分配问题,具有一定的理论价值。

关键词 三支决策理论,粗糙集,属性,权重,决策
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.8.054

Methodology of Attribute Weights Acquisition Based on Three-way Decision Theory

XUE Zhan-ao ZHU Tai-long XUE Tian-yu LIU Jie
(College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)
(Engineering Laboratory of Henan Province for Intelligence Business & Internet of Things, Xinxiang 453007, China)

Abstract Aiming at the problems of subjectivity in weight rules and uncertainty in the parameter value through traditional decision-making process, the methodology of attribute weights acquisition was studied based on the rough set and three-way decision theory. Two definitions of the attribute confirmation and attribute reduction were redefined respectively, a new method of attribute weights aquisition was presented and also compared with other methods by example, and then the efficiency of the presented method was demonstrated. This method is objective evaluation of attribute, which is based on data and without prior information. Using this method, decision maker can make decision practically and get the more reasonable weight distribution. This thesis has some value of theory about the study of attribute weight.

Keywords Three-way decision theory, Rough set, Attribute, Weight, Decision

1 引言

粗糙集理论(Rough Set)是波兰数学家 Pawlak[1] 于 1982 年提出的一种处理不确定性信息的理论方法,其理论研究和实际应用都取得了重要的成果。

经典的 Pawlak 粗糙集理论是将论域划分为 3 个区域:正域、负域、边界域,该理论不需要提供额外的预处理或者先验信息,能客观地对问题进行描述。2010 年,加拿大里贾纳大学的 Yao[2,3] 等人在决策粗糙集理论的基础上提出了三支决策理论,该理论是对传统的二支决策理论进行发展,即在接受和拒绝两种选择的基础上引入了不承诺决策(延迟决策)。正域中获取的事件对象用来选择接受,负域中获取的事件对象用来选择拒绝,而边界域中的事件对象用来选择延迟处理,即不能根据现有条件选择接受或者拒绝,需要进一步讨论,延迟对事件的决策,从而避免了强制选择接受或者拒绝带来的损失。

近期三支决策理论引起了一些学者的关注,成为了研究热点。姚静涛等[4] 将三支决策与博弈粗糙集结合起来进行研究;张宁等[5] 提出了基于 F-粗糙集的三支决策模型,用于解决多人决策时整体和局部的合理决策问题;商琳[6] 等提出了基于三支决策粗糙集的视频中异常行为检测方法;田海龙[7] 等将三支决策用于中文微博观点的句子识别研究;贾修一[8] 等提出了一种求三支决策阈值的模拟退火算法。文献[2,9]讨论了 Pawlak 粗糙集与二支决策和三支决策的关系,二支决策是三支或者多支决策的最终目标,三支决策归根到底要转化为二支决策[10]。

在决策问题的处理过程中,属性权重的确定是一个关键问题。权重反映了不同属性在决策过程中所起的作用,不同的权重会影响对问题的最终决策。而目前常用的权重确定方法有:层次分析法(Analytic Hierarchy Process, AHP)[11]、灰色关联分析法[12]、朴素贝叶斯方法[13]、主成分分析法(Princi-

到稿日期:2014-09-10 返修日期:2014-11-19 本文受国家自然科学基金计划项目(61273018),河南省基础与前沿技术研究计划项目(132300410174),河南省教育厅计划项目(14A520082),新乡市重点科技攻关计划项目(ZG14020)资助。
薛占熬(1963-),男,博士,教授,主要研究方向为人工智能基础理论,E-mail:xuezhanao@163.com;朱泰隆(1990-),男,硕士生,主要研究方向为博弈论、决策论和粗糙集理论;薛天宇(1991-),男,硕士生,主要研究方向为代数和机器学习;刘 杰(1989-),女,硕士生,主要研究方向为决策理论和模糊粗糙集理论。

pal Component Analysis,PCA)[14]。这些方法一般由专家根据主观性经验给出权重,其知识水平的不同和对于特定属性的偏好,会影响决策结果的客观性,会额外增加评价代价和分析的时间、空间复杂度。由于粗糙集不需要数据预处理和先验信息,有学者将粗糙集与决策问题权重赋值结合起来进行研究。刘盾等[15]通过定义两种属性重要度,提出了一种基于粗糙集理论中的属性重要度的权重构造方法,这对三支决策理论的条件属性权重构造研究具有重要意义。本文在粗糙集和三支决策理论的基础上,既考虑正域对于决策的贡献程度,又考虑负域对于决策的贡献程度,然后对 Pawlak 属性重要度进行改进并给出其新的定义,给出权重系数的生成规则,重新定义属性确定度和属性约简度,提出一种新的属性权重构造方法,通过实例分析比较,讨论该方法与其它权重方法的关系,阐明该方法的有效性,为三支决策提供了一种新的方法。

2 基本概念

2.1 粗糙集

粗糙集理论中的知识表达方式一般采用信息表或信息系统的形式,它是用四元有序组来表示的。

定义 1(信息系统)[16] 设四元有序组 $K=(U,A,V,d)$,其中 U 是对象的全体,即论域;A 是属性的全体;$V=\bigcup_{d\in A}V_a$,V_a 是属性的值域;$d:U\times A\to V$ 是一个信息函数,$d:A\to V$,$x\in U$,反映了对象 x 在 K 中的完全信息,其中 $d_x(a)=d(x,a)$。

定义 2[1] 给定信息系统 $K=(U,A,V,d)$,则对于任意 $X\subseteq U$ 和 U 上一个等价关系 R,每个子集 X 关于知识 R 的下近似和上近似分别为:

$$\underline{R}(X)=\bigcup\{Y|(\forall Y\in U/R)\wedge Y\subseteq X\}$$

$$\overline{R}(X)=\bigcup\{Y|(Y\in U/R)\wedge Y\cap X\neq\emptyset\}$$

$bnd_R(X)=\overline{R}(X)-\underline{R}(X)$称为 X 的 R 边界域;$pos_R(X)=\underline{R}(X)$称为 X 的 R 正域;$neg_R(X)=U-\overline{R}(X)$称为 X 的 R 负域。

定义 3(Pawlak 属性重要度)[17] 给定一个信息系统,$IS=(U,A,V,f)$,$\forall B\subseteq C$ 以及 $\forall a\subseteq C-B$:

(1)当 $pos_{IND(B\cup\{a\})}(C)=pos_{IND(B)}(C)$时,称属性 a 为属性集 B 的 C 不必要属性;

(2)当 $pos_{IND(B\cup\{a\})}(C)\neq pos_{IND(B)}(C)$时,称属性 a 为属性集 B 的 C 必要属性。

属性 a 为属性集 B 的 C 重要度为

$$\theta(a)=sig(a,B;C)=\frac{|pos_{IND(B\cup\{a\})}(C)|-|pos_{IND(B)}(C)|}{|U|}$$

2.2 三支决策

根据文献[3],实现三支决策首先需要引入实体的评价函数(决策函数),评价函数值称为决策状态值,其大小反映实体的好坏程度。其次需要引入阈值,根据阈值和决策状态值将论域中事件对象划分到正域、负域和边界域中,然后构造出相应的三支决策规则。对落在正域、负域和边界域的事件对象,分别选择接受、拒绝和不承诺决策(延迟决策),这就是所谓的三支决策。

定义 4[3] 决策粗糙集通过引入一对阈值 α 和 β 来定义正域、负域和边界域中的事件对象,设 $0\leqslant\beta<\alpha\leqslant1$,则 (α,β)-正域、边界域和负域可定义为:

$$pos_{(\alpha,\beta)}(X)=\{x\in U|P(X|[x])\geqslant\alpha\}$$

$$bnd_{(\alpha,\beta)}(X)=\{x\in U|\beta<P(X|[x])<\alpha\}$$

$$neg_{(\alpha,\beta)}(X)=\{x\in U|P(X|[x])\leqslant\beta\}$$

当 $\alpha=1,\beta=0$ 时,上面 3 个式子就转化为 Pawlak 粗糙集模型;当 $\alpha=\beta=0.5$ 时,其转化为 0.5-概率粗糙集模型;当 $\beta=1-\alpha$ 时,其转化为对称变精度概率粗糙集模型;当 $\beta\neq1-\alpha$ 时,其转化为非对称变精度概率粗糙集模型。

3 基于粗糙集与三支决策理论的权重构造方法

根据三支决策理论,决策者可以将事件划分为确定的类(接受或者拒绝选项)和不确定的类(延迟决策选项)。在实际生活中,对于直接判定接受或者拒绝的确定性事件,决策者能轻松地根据判断做出决策;而对于不确定事件,人们希望通过一定方法将其修正并判断为确定性事件,继而做出决策。

根据定义 2,传统的 Pawlak 属性重要度是从正域出发定义的,没有考虑负域中元素对属性重要程度的贡献。正域与负域中的元素均为确定的直接决策的类,而决策者重点关心划分到边界域中的不确定元素,它是需要延迟决策的。本文同时考虑了正域与负域中的元素贡献程度,定义了一种新的属性确定度,与传统 Pawlak 属性重要度相比,该属性确定度既考虑了正域中元素的贡献程度,又考虑了负域中元素的贡献程度,其决策更加客观合理。

定义 5(属性确定度) 给定一个信息系统,$IS=(U,A,V,f)$,$\forall B\subseteq C$ 且 $\forall a\subseteq C-B$,则属性 a 的属性确定度为:

$$\begin{aligned}&\xi(a)\\&=\left|\frac{|pos_{IND(B\cup\{a\})}(C)\cup neg_{IND(B\cup\{a\})}(C)|-|pos_{IND(B)}(C)\cup neg_{IND(B)}(C)|}{|U|}\right|\\&=\left|\left(1-\frac{|(\overline{R}_{(B\cup\{a\})}(C)-\underline{R}_{(B\cup\{a\})}(C))|}{|U|}\right)-\left(1-\frac{|(\overline{R}_{(B)}(C)-\underline{R}_{(B)}(C))|}{|U|}\right)\right|\\&=\left|\frac{|(\overline{R}_{(B\cup\{a\})}(C)-\underline{R}_{(B\cup\{a\})}(C))|}{|U|}-\frac{|(\overline{R}_{(B)}(C)-\underline{R}_{(B)}(C))|}{|U|}\right|\end{aligned}$$

定义 3 的 Pawlak 属性重要度与定义 5 中的属性确定度只能判断必要属性的重要度,忽略了某些非必要属性在约简中的贡献度。因此在文献[16]中的 ξ-重要度基于属性约简集定义的启发下,重新定义属性约简度,以弥补某些非必要属性在约简中的贡献度无法体现的不足。

定义 6(属性约简度) 设 U 为论域,共有 n 个属性,m 个约简,$F(x,f_k)$表示 U 中所有约简 f_k 中属性个数为 x 的约简集个数,C_n^x 表示从 n 个元素中取 x 的组合数,即属性 a 的属性约简度为:

$$\rho(a)=\frac{\sum_{x=1}^{n}\sum_{k=1}^{m}F(x,f_k)}{\sum_{x=1}^{n}C_n^x}$$

该属性约简度与文献[16]中 ξ-属性重要度相比,两者都是通过约简的角度来衡量属性的重要程度,但本文提出的属性约简度是从宏观、整体来考虑,通过对整个约简集合的相同属性集个数与其组合数求比值来进行计算的,而文献[16]中 ξ-属性重要度是通过单独的属性出现在约简集中的个数与整个约简集个数求比值来计算的。本文中的属性约简度能较全面客观地反映某属性对整体的贡献程度。

属性确定度和属性约简度均是在三支决策理论框架下定义的,没有主观条件的干扰和先验信息的影响,定义 5 弥补了某些非必要属性在约简中的贡献度无法体现的不足,定义 6

是从宏观上对整体约简集合进行评价分析。综合定义5和定义6中的两种度量标准，本文给出一种新的权重构造方法。某属性 a 的权重构造方法具体步骤如下。

(1)根据定义5，计算属性 a 的属性确定度 $\xi(a)$。

(2)根据定义6，计算属性 a 的属性约简度 $\rho(a)$。

(3)计算权重 $\omega(a)=\lambda\xi(a)+(1-\lambda)\rho(a)$。

其中，λ 为定义的一个权重参数，传统的权重参数一般由专家人为给出，文献[16]认为 λ 为常数，未给出计算公式，本文基于数据集本身给出参数 λ 的定义：$\lambda=\left|\frac{N(core)}{N(U)}-0.5\right|$，$N(core)$ 为核属性个数，$N(U)$ 为全部属性个数；当不存在核属性时，$\frac{N(core)}{N(U)}=0$；当全部属性均为核属性时，$\frac{N(core)}{N(U)}=1$；λ 取 $\frac{N(core)}{N(U)}$ 与0.5作差的绝对值，来保证 $\lambda\neq0$。

(4)归一化处理。

4 实例分析

本节在上述理论的基础上，选用UCI数据库中皮马印第安人糖尿病数据集(Pima Indians Diabetes Data Set)对该权重构造方法进行实例分析，其原始数据网址为 http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes，共768条数据，包括怀孕次数(Number of times pregnant)、口服葡萄糖2小时血糖浓度(Plasma glucose concentration a 2-hours in an oral glucose tolerance test)、舒张压(Diastoic blood pressure (mm Hg))、肱三头肌皮肤褶皱厚度(Triceps skin fold thickness(mm))、2小时血清胰岛素(2-Hour serum insulin(mu U/ml))、身体质量指数(Body mass index)、糖尿病血统函数(Diabetes pedigree function)、年龄(Age(years))等8项条件属性和决策属性 M(类变量(Class variable))，使用Excel提供的随机函数选取其中9条数据(数据排序号分别为130、608、332、34、531、99、444、540、448，分别对应 $u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8,u_9$)来分析讨论。$x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8$ 表示8种条件属性，M 表示决策属性，离散化后条件属性的值域划分为 A,B,C,D 4个等级，决策属性值域为{0,1}，0和1分别表示没有患病者和患病者，如表1所列。

表1 皮马印第安人糖尿病情况随机抽样数据的决策表

数据号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	M
u_1	0	105	84	0	0	27.9	0.741	62	1
u_2	1	92	62	25	41	19.5	0.482	25	0
u_3	2	87	58	16	52	32.7	0.166	25	0
u_4	6	92	92	0	0	19.9	0.188	28	0
u_5	2	122	60	18	106	29.8	0.717	22	0
u_6	6	93	50	30	64	28.7	0.356	23	0
u_7	8	108	70	0	0	30.5	0.955	33	1
u_8	3	129	92	49	155	36.4	0.968	32	1
u_9	0	95	80	45	92	36.5	0.33	26	0

对表1中的每一列的属性值进行离散化，得到表2，记 $U=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8,u_9\}$，$S=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$。通过波兰Warsaw大学研制的Rosetta软件对表2中的数据进行计算，得整个决策表的约简集合如下：

$R_1=\{x_1,x_4\}$，$R_2=\{x_4,x_7\}$，$R_3=\{x_2,x_4\}$，

$R_4=\{x_2,x_7\}$，$R_5=\{x_7,x_8\}$，$R_6=\{x_2,x_3\}$，

$R_7=\{x_3,x_5\}$，$R_8=\{x_3,x_4\}$，$R_9=\{x_3,x_6\}$，

$R_{10}=\{x_2,x_6\}$，$R_{11}=\{x_1,x_7\}$，$R_{12}=\{x_3,x_7\}$，

$R_{13}=\{x_5,x_7\}$，$R_{14}=\{x_2,x_8\}$，$R_{15}=\{x_5,x_6\}$，

$R_{16}=\{x_1,x_3,x_8\}$，$R_{17}=\{x_1,x_5,x_8\}$

表2 皮马印第安人糖尿病情况随机抽样数据的赋值决策表

数据号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	M
u_1	A	C	C	A	A	B	C	C	1
u_2	A	B	B	C	B	A	B	B	0
u_3	B	A	B	B	B	C	A	B	0
u_4	C	B	D	A	A	A	A	B	0
u_5	B	D	B	B	D	B	C	A	0
u_6	C	B	A	C	C	B	B	A	0
u_7	D	C	C	A	A	B	D	B	1
u_8	B	D	D	D	D	C	D	B	1
u_9	A	B	C	D	C	C	B	B	0

通过观察可知决策表没有核属性，所以 $N(core)=0$，$\lambda=0.5$，利用本文给出的权重确定方式，则8个条件属性的最终权重分别为：

$w_1=0.045$，$w_2=0.227$，$w_3=0.091$，$w_4=0.182$

$w_5=0.045$，$w_6=0.136$，$w_7=0.273$，$w_8=0.061$

选取层次分析法(Analytic Hierarchy Process，AHP)、主成分分析法(Principal Component Analysis，PCA)与本文提出的方法进行比较。层次分析法(AHP)将与决策相关的元素分解为多个层次进行分析。算例中8个条件属性被分为3个层次，如图1所示，分别为决策目标层、中间层、方案层。其中决策目标层中只包含皮马印第安人糖尿病情况(Pima Indians Diabetes Data)，对应决策系统的决策目标；中间层包括生理特征(physiological feature)、身体参数(bodyparameters)、综合数据(synthetic data)；方案层将中间层继续进行划分，年龄(Age)和怀孕次数(Number of times pregnant)归属生理特征(physiological feature)，口服葡萄糖2小时血糖浓度(Plasma glucose concentration a 2 hours in an oral glucose tolerance test)、舒张压(Diastolic blood pressure)、2小时血清胰岛素(2-Hour serum insulin)和肱三头肌皮肤褶皱厚度(Triceps skin fold thickness)归属身体参数(body parameters)，身体质量指数(Bodymass index)和糖尿病血统函数(Diabetes pedigree function)归属综合数据(synthetic data)。

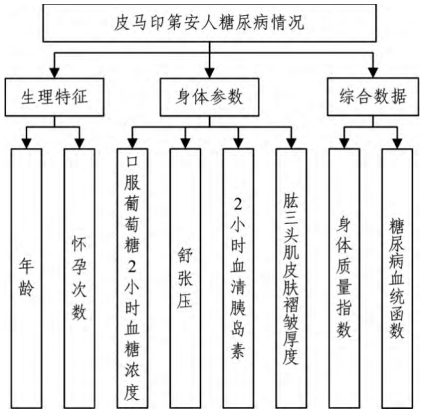

图1 层次分析法分层示意图

根据相关医学系统专家的意见，构造层次分析矩阵如下：

中间层 $\begin{pmatrix}1&1/3&1/3\\3&1&1/2\\3&2&1\end{pmatrix}$

方案层 $\begin{pmatrix}1&1/2\\2&1\end{pmatrix}$，$\begin{pmatrix}1&1/3&1/2&1/5\\3&1&1/3&1/3\\2&3&1&1/2\\5&3&2&1\end{pmatrix}$

通过 Yaahp 软件采用规范列平均法(和法)对矩阵进行计算,计算步骤如下:

(1)将矩阵每一列向量归一化得 $\widetilde{w}_{ij}=\frac{a_{ij}}{\sum_{i=1}^{n}a_{ij}}$;

(2)对 $\widetilde{w}_{ij}$ 按行求和得 $\widetilde{w}_i=\sum_{j=1}^{n}\widetilde{w}_{ij}$;

(3)归一化,$w_i=\frac{\widetilde{w}_i}{\sum_{i=1}^{n}\widetilde{w}_i}$,得 $W=(w_1,w_2,\cdots,w_n)^{\mathrm{T}}$,即权重向量。由此得到最终的 8 个条件属性权重为:

$w_1=0.0944$,$w_2=0.0305$,$w_3=0.0511$,$w_4=0.1566$,

$w_5=0.0916$,$w_6=0.1749$,$w_7=0.3498$,$w_8=0.0472$

主成分分析法(PCA)是通过将多个变量进行变换,从中筛选出代表性变量的分析方法。使用 Spss 软件求得解释的总方差和成分矩阵,以主成分的方差贡献率为权重,对该指标在各主成分线性组合中的系数加权平均归一化,求得指标在不同主成分线性组合中的系数。然后对指标分布在两个主成分内的系数做加权平均,最后对其归一化处理,得到的 8 个条件属性的权重值分别为:

$w_1=0.053$,$w_2=0.318$,$w_3=0.082$,$w_4=0.072$,

$w_5=0.189$,$w_6=0.075$,$w_7=0.302$,$w_8=-0.090$

将 3 种方法得到的权重进行比较的结果如下:

AHP 方法:$w_7>w_6>w_4>w_1>w_5>w_3>w_8>w_2$;

PCA 方法:$w_2>w_7>w_5>w_3>w_6>w_4>w_1>w_8$;

本文方法:$w_7>w_2>w_4>w_6>w_3>w_8>w_1\geqslant w_5$。

对比结果如表 3 和图 2 所示。层次分析法(AHP)得到的权重结果,依赖于判别矩阵和专家的主观判断,同时过分强调属性 7 糖尿病血统函数(Diabetes pedigree function)的重要性;主成分分析法(PCA)由于需提取代表性主成分,增加了额外的先验信息和评价标准,且出现了不合实际的负值权重,因此不能真实客观地反映权重的分布;本文提出的方法,不同属性权重有明显区分,整体分布合理,属性 2 和属性 7 均为医学仪器检测数据,其权重较高,能较客观地反映患者身体的实际情况。

表 3　3 种方法权重的比较

权重方法	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
AHP	0.0944	0.0305	0.0511	0.1566	0.0916	0.1749	0.3498	0.0472
PCA	0.053	0.318	0.082	0.072	0.189	0.075	0.302	−0.090
本文方法	0.045	0.227	0.091	0.182	0.045	0.136	0.273	0.061

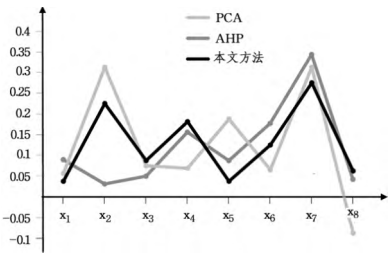

图 2　3 种属性权重方法结果的对比

结束语　本文在三支决策理论基础上对属性权重构造方法进行研究,重新定义了属性确定度和属性约简度,提出了一种基于三支决策理论的属性权重构造方法,通过 UCI 数据库中皮马印第安人糖尿病数据集的例子进行验证,分析过程是基于信息本身,没有任何先验信息,完全依靠数据。最后通过实例对比分析了层次分析法、主成分分析法与本文提出的方法,验证了本文方法的有效性。决策者用此方法可以客观地处理决策问题,做出比较符合实际的决策。

参 考 文 献

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

[2] 刘盾,姚一豫,李天瑞. 三支决策粗糙集[J]. 计算机科学,2011,38(1):246-250
Liu Dun, Yao Yi-yu, Li Tian-rui. Three-way Decision-theoretic Rough Sets[J]. Computer Science,2011,38(1):246-250

[3] 刘盾,李天瑞,苗夺谦,等. 三支决策与粒计算[M]. 北京:科学出版社,2013
Liu Dun, Li Tian-rui, Miao Duo-qian, et al. Three-way Decision and Granular Computing[M]. Beijing: Science Press,2013

[4] 贾修一,商琳,周献中,等. 三支决策理论与应用[M]. 南京:南京大学出版社,2012
Jia Xiu-yi, Shang Lin, Zhou Xian-zhong, et al. Three decision-making theory and application[M]. Nanjing: Nanjing University Press,2012

[5] 张宁,邓大勇,裴明华. 基于 F-粗糙集的三支决策模型[J]. 南京大学学报(自然科学版),2013,49(5):582-587
Zhang Ning, Deng Da-yong, Pei Ming-hua. A model of three-way decision based on F-rough sets[J]. Journal of Nanjing University(Natural Science),2013,49(5):582-587

[6] 谢骋,商琳. 基于三支决策粗糙集的视频异常行为检测[J]. 南京大学学报(自然科学版),2013,49(4):475-482
Xie Cheng, Shang Lin. Detection of abnormal behavior in video using three-way decision rough sets[J]. Journal of Nanjing University(Natural Science),2013,49(4):475-482

[7] 田海龙,朱艳辉,梁韬,等. 基于三支决策的中文微博观点句识别研究[J]. 山东大学学报(理学版),2014,49(8):58-65
Tian Hai-long, Zhu Yan-hui, Liang Tao, et al. Research on identificating Chinese micro-blog opinion sentence based on three-way decisions[J]. Journal of Shangdong University(Natural Science),2014,49(8):58-65

[8] 贾修一,商琳. 一种求三支决策阈值的模拟退火算法[J]. 小型微型计算机系统,2013,34(11):2603-2606
Jia Xiu-yi, Shang Lin. A Simulated Annealing Algorithm for Learning Thresholds in Three-way Decision-theoretic Rough Set Model[J]. Journal of Chinese Computer Systems,2013,34(11):2603-2606

[9] Yao Yi-yu. The superiority of three-way decision in probabilistic rough set models[J]. Information Sciences,2011,181:1080-1096

[10] 姚一豫. 三支决策研究的若干问题[M]//刘盾,李天瑞,苗夺谦,等. 三支决策与粒计算. 北京:科学出版社,2013:1-13
Yao Yi-yu. Some problems concerning the study of three-way decision [M] // Liu Dun, Li Tian-rui, Miao Duo-qian, et al. Three-way Decision and Granular Computing. Beijing: Science Press,2013:1-13

[11] 邓雪,李家铭,曾浩健,等. 层次分析法权重计算方法分析及其应用研究[J]. 数学的实践与认识,2012,42(7):93-100
Deng Xue, Li Jia-ming, Zeng Hao-jian, et al. A computational weight analysis ofthe method based on AHP and application study[J]. Mathematics in Practice an Theory,2012,42(7):93-100

(下转第 272 页)

性能的影响，适合表达复杂细微差异化网络数据的非线性和随机散布性语义特征；采用本文算法，通过 Dopplerlet 变换匹配投影，计算投影后的残差信号，对复杂细微差异化网络数据的语义特征提取结果准确，语义表达能力提高明显，能有效区分差异网络数据中的冗余数据和残差数据。图 4 的语义特征表达结果直观地展示了本文算法的优越性能。

为定量分析本文算法的特征提取性能，通过 1000 次 Monte Carlo 实验，在不同的干扰数据信噪比环境下，对网络数据进行语义特征检测的性能分析，得到如图 5 所示的结果，其中，FRFT 表示傅里叶变换，FRFT-FOMCS 表示分数阶傅里叶变换。从图 5 可知，89.5%以上的语义的相似度值大于 0.1，随着信噪比的增大，查准率不断下降，当阈值为 0.9 时，查准率为98.7%，采用本文算法能有效提高对复杂细微差异化网络数据的检测识别性能，结果优越于传统算法。本文算法对差异化网络数据的语义关系表达清晰，性能优越。

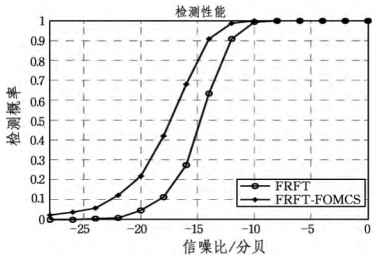

图 5　语义特征提取检测性能对比

结束语　对复杂细微差异化网络数据的语义特征提取和语义表达是实现网络数据挖掘和智能分析的基础，是实现 Web 网络数据准确识别和检索的关键。研究网络数据的语义特征提取算法，在数据挖掘和网络 Web 数据库访问等领域意义重大。复杂细微差异化网络数据的语义特征具有非线性和随机散布性，其主题分布广、更新频率大，提取困难。本文提出一种基于 Dopplerlet 变换匹配投影的网络数据特征的语义优化提取算法。利用 Dopplerlet 变换匹配投影的自相似特性以及能自适应匹配语义的非线性谱特征的特点，进行算法改进。研究结果表明，改进算法能克服中心频率变化对算法性能的影响，明显提高语义表达能力，能有效区分差异网络数据中的冗余数据和残差数据，提高对复杂细微差异化网络数据的检测识别和检索能力。

参 考 文 献

[1] Bimal K M, Gholam M A. Differential epidemic model of virus and worms in computer network [J]. International Journal of Network Security, 2012, 14(3): 149-155

[2] Zhu Q Y, Yang X F, Yang L X, et al. Optimal control of computer virus under a delayed model [J]. Applied Mathematics and Computation, 2012, 218(23): 11613-11619

[3] Zhu Q Y, Yang X F, Ren J, et al. Modeling and analysis of the spread of computer virus[J]. Communications in Nonlinear Science and Numerical Simulation, 2012, 17(12): 5117-5124

[4] Zahran B M, Kanaan G. Text feature selection using particle swarm algorithm [J]. World Applied Sciences Journal, 2009, 25(7): 69-74

[5] 姜大庆，周勇，夏士雄. 基于语义描述与优化的网络性能数据聚类方法[J]. 计算机应用，2012，32(6)：1522-1525
Jiang Da-qing, Zhou Yong, Xia Shi-xiong. Network performance data clustering method based on semantic description and optimization[J]. Journal of Compater Applications, 2012, 32(6): 1522-1525

[6] Hu W M, Hu M, Maybank S. Adaboost based algorithm for network intrusion detection [J]. IEEE Transactions on Systems, Man and Cybernetic, Parb B: Cybernetics, 2008, 38(2): 577-583

[7] 何永强，谷春英. 基于子任务区域分片下的分布式空间查询处理与并行调度方法[J]. 科技通报，2014，30(1)：110-116
He Yong-qiang, Gu Chun-ying. Distributed Spatial Query Processing and Parallel Schedule Based on Zonal Fragmentation[J]. Bulletin of Science and Technology, 2014, 30(1): 110-116

[8] 邓兵，陶然，平殿发，等. 基于分数阶傅里叶变换补偿多普勒徙动的动目标检测算法[J]. 兵工学报，2009，30(10)：1034-1039
Deng Bing, Tao Ran, Ping Dian-fa, et al. Moving-Target-Detection Algorithm with Compensation for Doppler Migration Based on FRFT[J]. Acta Armanentarii, 2009, 30(10): 1034-1039

[9] 蒋芸，陈娜，明利特，等. 基于 Bagging 的概率神经网络集成分类算法[J]. 计算机科学，2013，40(5)：242-246
Jiang Yun, Chen Na, Ming Li-te, et al. Bagging-based probability Neural Network Ensemble Classification Algorithm[J]. Computer Science, 2013, 40(5): 242-246

[10] 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
Chen Hao, Yang Jun-an, Zhuang Zhen-quan. The Core of Attributes and Minimal Attributes reduction in Variable Precision Rough Set[J]. Chinese Journal of Computers, 2012, 35(5): 1011-1017

(上接第 268 页)

[12] 钱玲飞，杨建林，张莉. 基于灰色关联度的学科创新力影响因素权重分析——以情报学为例[J]. 图书情报工作，2011，55(16)：37-40
Qian Ling-fei, Yang Jian-lin, Zhang Li. Weight Analysis on Influence Factors of Disciplinary Creativity based on Grey Relation-Take Information Science as an Example[J]. Library and Information Service, 2011, 55(16): 37-40

[13] 张明卫，王波，张斌，等. 基于相关系数的加权朴素贝叶斯分类算法[J]. 东北大学学报(自然科学版)，2008，29(7)：952-955
Zhang Ming-wei, Wang Bo, Zhang Bin, et al. Weighted Naive Bayes Classification Algorithm Based on Correlation Coefficients [J]. Journal of Northeastern University (Natural Science), 2008, 29(7): 952-955

[14] 韩小孩，张耀辉，孙福军，等. 基于主成分分析的指标权重确定方法[J]. 四川兵工学报，2012，33(10)：124-126
Han Xiao-hai, Zhang Yao-hui, Sun Fu-jun, et al. A method of determining index weights based on PCA[J]. Journal of Sichuan Ordnance, 2012, 33(10): 124-126

[15] 刘盾，胡培，蒋朝哲. 一种基于粗糙集理论的属性权重构造方法[J]. 系统工程与电子技术，2008，30(8)：1482-1484
Liu Dun, Hu Pei, Jiang Chao-zhe. New methodology of attribute weights acquisition based on rough sets theory[J]. Systems Engineering and Electronics, 2008, 30(8): 1482-1484

[16] 张文修，吴伟志. 粗糙集理论介绍和研究综述[J]. 模糊系统与数学，2000，14(4)：1-12
Zhang Wen-xiu, Wu Wei-zhi. An Introduction and a Survey for the Studies of Rough Set Theory[J]. Fuzzy Systems and Mathematics, 2000, 14(4): 1-12

[17] 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
Wang Guo-ying. Rough set theory and knowledge acquisition [M]. Xi'an: Xi'an Jiaotong University Press, 2001