基于粗糙集的不完备谣言信息系统的知识获取与决策

王 标[1], 卫红权[1,2], 王 凯[1,2], 刘树新[1,2], 江昊聪[1,2]

[1]中国人民解放军战略支援部队信息工程大学 郑州 中国 450001
[2]国家数字交换系统工程技术研究中心 郑州 中国 450002

摘要 网络谣言可能扰乱人们的思想、心理和行为, 引发社会震荡、危害公共安全, 而微博等社交平台的广泛应用使得谣言造成的影响与危害变得更大, 因此, 谣言检测对于网络空间的有序健康发展具有重要的意义。当前谣言的自动检测技术更多关注检测模型的构建和输入数据的表现形式, 而在改善数据质量以提高谣言识别效果方面的研究很少。基于此, 本文将粗糙集理论应用于不完备谣言信息系统进行知识获取与决策, 实质上是通过粗糙集理论解决不完备谣言信息系统的不确定性度量, 冗余性以及不完备性等问题, 以获得高质量的数据, 改善谣言检测效果。首先系统总结了粗糙集理论中不确定性度量的方法, 包括香农熵、粗糙熵、Liang 熵以及信息粒度等四种不确定度量方法, 并整理和推导了这四种不确定度量方法从完备信息系统到不完备信息系统的一致性拓展。基于上述总结的四种不确定度量方法, 提出了基于最大相关最小冗余(MCMR, Maximum Correlation Minimum Redundancy)的知识约简算法。该方法基于熵度量方式, 能够综合考量决策信息与冗余噪音, 在 UCI 及 Weibo 等 8 个数据集上实验验证, 结果表明本文算法优于几种基线算法, 能够有效解决信息系统的冗余性。另外, 提出了一种基于极大相容块的不完备决策树算法, 在不同缺失程度数据上实验验证, 结果表明本文算法能够有效解决信息系统的不完备性。

关键词 谣言检测; 粗糙集; 不完备信息系统; 最大相关最小冗余; 极大相容块
中图法分类号 TP391.1; O236 DOI 号 10.19363/J.cnki.cn10-1380/tn.2024.03.02

Knowledge Acquisition and Decision Making in Incomplete Rumor Information System based on Rough Set

WANG Biao[1], WEI Hongquan[1,2], WANG Kai[1,2], LIU Shuxin[1,2], JIANG Haocong[1,2]

[1]PLA Strategic Support Force Information Engineering University, Zhengzhou 450002, China
[2]National Digital Switching System Engineering and Technological R&D Center, Zhengzhou 450002, China

Abstract Online rumors may disrupt people's thoughts, psychology and behavior, cause social shocks and endanger public safety. The widespread use of social platforms such as Weibo makes the impact and harm caused by rumors even greater. Therefore, rumor detection is of great significance to the orderly and healthy development of cyberspace. The current automatic detection techniques for rumors focus more on the construction of detection models and the representation of input data, while there is little research on improving the quality of data to improve the effect of rumor detection. Based on this idea, this paper applies the rough set theory to the incomplete rumor information system for knowledge acquisition and decision-making. In essence, to obtain high-quality data and improve rumor detection, the rough set theory is used to solve the uncertainty measurement, redundancy, and incompleteness of the incomplete rumor information system. Firstly, it systematically summarizes the methods of uncertainty measurement in rough set theory, including four uncertainty measurement methods such as Shannon entropy, rough entropy, Liang entropy, and information granularity, and organizes and derives the consistent expansion of the four uncertainty measurement methods from complete information system to incomplete information system. Based on the four uncertainty measurement methods summarized above, a knowledge reduction algorithm based on Maximum Correlation Minimum Redundancy (MCMR) is proposed. The method is based on entropy measurement, which can comprehensively consider decision information and redundant noise. Experiments on 8 data sets such as UCI and Weibo show that the algorithm in this paper is superior to several baseline algorithms and can effectively solve the redundancy of the information system. In addition, this paper proposes an incomplete decision tree algorithm based on maximal consistent blocks. Experiments on data with different degrees of missingness show that the algorithm in this paper can effectively solve the incompleteness of the information system.

通讯作者: 卫红权, 博士, 研究员, Email: whq@ndsc.com.cn。
本课题得到中原英才计划项目(No. 212101510002)资助。
收稿日期: 2022-05-24; 修改日期: 2022-08-19; 定稿日期: 2023-11-01

Key words rumor detection; rough set; incomplete information system; maximum correlation minimum redundancy; maximal consistent blocks

1 引言

微博等社交平台的广泛应用缩短了信息传播的周期, 扩大了信息传播的范围, 使得虚假信息传播更加容易实现, 造成的影响和危害变得更大[1]。比如当前新冠疫情传播的谣言四处而起, 给政府、社会和人们造成了巨大的困扰。

越来越多的现实网络被人们研究, 从建模到信息挖掘[2], 社交网络[3]领域的谣言识别一直是近年来研究的热点。谣言识别本质上是一个分类问题, 从早期的传统机器学习方法到深度学习方法, 我们一直致力于寻找更好的分类模型, 提高谣言识别的效果。早期自动检测大多基于传统的机器学习方法, 如决策树[4-5]、支持向量机[4-8]、随机森林[5, 9]、逻辑回归[9]、朴素贝叶斯[9]、传统自然语言处理[10]等方法, 采用手工提取特征, 耗时耗力, 而且不能提取深层的特征。近年来的自动检测方法越来越多地采用深度学习方法, 如卷积神经网络(CNN)[11-12], BERT 预训练模型[13]、长短期记忆网络(LSTM)[12, 14], RNN[15]、门控循环单元 GRU[16-17], 图卷积神经网络(GCN)[18-19]等。深度学习模型能够自动提取特征, 且可以进行深度拟合, 识别谣言效果好, 但是可解释性不强, 鲁棒性弱, 调参困难等问题也不容忽视。

同时, 当前的文献更注重从模型构建方面提高谣言的识别效果, 从数据分析与处理方面提高数据质量的文献较少。而数据的复杂性[20]、不确定性[21]、不完备性[22]和冗余性[23]是在网络空间发展中逐渐具备的特征。

数据的质量将直接影响到模型的性能。当前数据的复杂性越来越高, 突出表现在数据的量级以及数据源的种类。本文所描述的数据复杂性是数据源种类中的数值多样性, 包括符号型、连续型、离散型、区间型、模糊型、标签类、文本以及空值等。混合类型数据不能直接用于模型的训练, 需要进一步处理, 而数据处理的好坏将直接影响到模型的性能。而数据的不确定性包括随机性、粗糙性以及模糊性, 分别是因为因果关系不确定、知识不充分以及定义不明确造成的。不同的不确定性数据需要不一样的模型处理, 反过来也会对模型的性能产生影响。同时, 在数据获取或处理时有些信息是缺失的, 即不完备性, 这部分数据可能具有极大价值, 但因为缺失常被直接删掉或用其他值代替, 没有有效地发挥其数据的作用和效果。还有, 相似冗余的数据, 在给我们带来可用信息的同时, 也造成了不必要的噪音以及计算消耗。沿着这个思路, 本文从如何解决谣言信息系统的不确定性, 不完备性以及冗余性角度出发进行谣言识别。

粗糙集理论是 Pawlak[24]提出的一种数学工具, 它能够定量分析处理不确定, 不一致, 不完备的信息与知识。本文将粗糙集理论应用于谣言识别, 在不完备谣言信息系统中进行知识获取与决策, 本质上是通过粗糙集理论解决谣言信息系统的不确定性度量, 冗余性以及不完备性。本文的贡献主要有:

- 将粗糙集理论应用于谣言检测, 系统总结了粗糙集理论中的不确定性度量方法和不同度量方法在完备信息系统和不完备信息系统中的一致性表达。
- 提出了一种基于最大相关最小冗余的知识约简算法, 该算法能够综合考量决策信息与冗余噪声, 实验结果表明, 算法优于多种基线算法, 能有效解决信息系统的冗余性。
- 基于极大相容块提出了一种以决策粗糙指标为核的不完备决策树模型, 可有效解决信息系统的不完备性。

2 相关工作

2.1 谣言检测最新研究

随着社交网络的不断发展以及谣言特征的不断变化, 谣言检测经历了人工事实核查、传统机器学习、深度学习等阶段。

目前的研究大部分都是基于深度学习方法, 而且多采用图神经网络(GNN)来提取谣言的传播特征。Bian 等人[18]首次将 GCN 应用于社交媒体的谣言检测, 通过自顶向下和自底向上的双向 GCN 提取谣言的传播和扩散特征。为了动态调整传播图中每个节点的权重, Wu 等人[25]采用基于注意力机制的图神经网络模型, 进一步提升了谣言检测的性能。然而当前的检测模型大多基于静态网络, 针对这个问题, Song 等人[26]提出了一种新的基于时间传播的动态检测框架, 可以融合结构、内容语义和时间信息。现有的模型大部分都是有监督学习模型, 需要大量的标签数据, 而数据的标注是一项费时费力的工程。为此, He 等人[27]引入了对比自监督学习来有效实现事件增强, 并缓解有限的标签数据问题。

但总的来讲, 深度学习模型适合进行谣言识别的感知与预警, 但其可解释性不足的问题让人们无法完全信服模型的判断, 而人工选择特征, 可解释性较强的传统机器学习算法更适合进行决策。

2.2 粗糙集理论研究与应用

粗糙集理论是 Pawlak 教授[24]于 1982 年提出的。后来, Ivo 等人[28], Theresa 等人[29], Liang 等人[30-31]将香农熵及其变体应用于粗糙集理论中测度系统的不确定性。Liang 等人[22, 32-33]结合补集提出了熵的新形式, 即 Liang 熵, 同时整理证明了 Liang 熵与知识粒度, 香农熵与粗糙熵的关系。

传统粗糙集的属性约简算法依赖于严格的等价关系, 有很多局限。众多学者在此基础上, 提出决策粗糙集[34], 概率粗糙集[35], 多粒度粗糙集[36], 变精度粗糙集[37]等, 并应用于不相容决策信息表[38], 不完备决策信息表[39], 连续型属性决策信息表[40], 有序型属性决策信息表[41], 属性动态变化决策信息表[42]等。

粗糙集理论广泛应用于各种领域, 如种类识别、物流模式决策、医疗诊断、智能识别、机器学习、数据分析以及网络检测等[43]。周正国[44]运用模糊粗糙集属性约简的方法, 在海量的数据挖掘中, 利用 Canopy 算法对 K-means 算法进行改进, 实现了 K-means 算法在 Hadoop 平台上的并行化计算。郭威[45]基于贝叶斯粗糙集提出了大数据频繁项挖掘方法, 具有较高的鲁棒性, 大大提高了数据挖掘的准确率和运行时间。

2.3 不完备信息系统知识获取研究

现实中, 不精确的数据测量造成的误差、不同的数据理解以及数据获取的严格限制等都可能导致信息系统的不完备性[46]。在粗糙集理论中, 知识被看作是一种分类的能力。不完备信息系统中的知识获取一般有两种方式: 间接处理与直接处理[46]。

间接处理, 就是将不完备信息系统进行完备化, 然后用传统的知识获取方法进行处理。常见的完备化操作有填充值与删除值。填充值包括填充均值、最大值、最小值、0、1 等, 删除值就是不考虑含有未知属性值的对象, 直接删除。

直接处理, 就是将粗糙集中的部分概念在不完备信息系统下进行适当拓展。主要采用了相似关系、相容关系和极大相容关系等。Slowiński 等人[47]利用粗糙集理论对对象属性值的不精确、不完备等进行了建模与描述, 对不确定数据进行了推理。Lipski 等人[48]基于相容关系提出了一种在信息不完整数据库中进行查询的模型。

3 粗糙集基础

3.1 完备信息系统

$S=(U,A,V,f)$是一个信息系统, 其中U为对象的非空有限集合, 称为论域(universe); A为属性的非空有限集合; 特别地, $C\cup D=A$, C为条件属性集合, D为决策属性集合, 带有决策属性集合D的信息系统又被称为决策信息系统。

$V=\bigcup_{a\in A}V_a$, V_a表示属性a的值域; $f:U\times A\to V$是一个函数, 即对$\forall x\in U,a\in A$, 有$f(x,a)\in V_a$。

$S=(U,A,V,f)$可简记为$S=(U,A)$。形式上, 用信息表来表示, 如表 1。它的行代表对象, 也是数据中的样本; 列代表属性, 属性值对应数据样本的不同维度的特征值。

在粗糙集理论中, 知识代表的是分类能力, 也就是利用对象的条件属性值区分其决策属性值的能力, 传统的粗糙集理论建立在严格的不可区分关系上[49]。令$B\subseteq A$, 定义属性集B的不可区分关系(indiscernibility) $IND(B)$为

$$IND(B)=\{(x,y)\in U\times U|\ \forall a\in B, f(x,a)=f(y,a)\}$$

如果$(x,y)\in IND(B)$, 则称x和y是B不可区分的。符号$U/IND(B)$表示不可区分关系$IND(B)$在U上导出的划分。符号$[x]_B$表示包含$x\in U$的B等价类。

3.2 不完备信息系统

设$S=(U,A,V,f)$是一个信息系统, 某些时候, 对一些对象而言, 一些属性值可能是缺损的。如果至少有一个属性$a\in A$, V_a中含有空值, 则称S是一个不完备信息系统(incomplete information system), 否则它是完备信息系统。这表明完备信息系统是不完备信息系统的特例。并用*表示空值[50]。

设$B\subseteq A$, 我们定义相容关系$SIM(B)=\{(u,v)\in U\times U|\ \forall a\in B, f(u,a)=f(v,a)\ or\ f(u,a)=*\ or\ f(v,a)=*\}$

易知$SIM(B)=\bigcap_{a\in B}SIM(\{a\})$, 令$S_B(u)$表示对象集$\{v\in U|(u,v)\in SIM(B)\}$。$S_B(u)$表示与$U$具有相容关系的所有对象集合, 相对$B$而言, 这些关系可能是可区分的, 也可能是不可区分的。

$U/SIM(B)=\{S_B(u)|\ u\in U\}$, $U/SIM(B)$中的元素称为相容类。$U/SIM(B)$中的相容类一般不构成U的划分, 它们构成U的覆盖, 即对于每一个

$u \in U$ 有 $S_B(U) \neq \varnothing$，且 $\bigcup_{u \in U} S_B(u) = U$。

3.3 极大相容块

相比于相容关系，极大相容块是对不完备信息划分的更简洁、更凝练的表示方式，下面给出极大相容块的定义。

定义 1[51]. 设 $S = (U, A, V, f)$ 是一个不完备信息系统，$B \subseteq A$，$X \subseteq U$，我们称 X 关于 B 是相容的，如果对任意的 $x, y \in X$ 有 $(x, y) \in SIM(B)$。如果不存在一个对象子集 $Y \subseteq U$ 使得 $X \subseteq Y$ 且 Y 关于 B 是相容的，则 X 被称为一个极大相容子集或极大相容块[52]。

极大相容块也是极大对象集合，里面的所有对象都是相似的或者相容的，是不可区分的。以 $C_{\text{block}}(B)$ 表示由 $B \subseteq A$ 导出的所有极大相容块所构成的类[52]，如例 1。

例 1. 表 1 是一个简化的不完备谣言决策信息表。其中 City, Province, Friends, Followers, Gender 是系统的条件属性, d 为决策属性。下面 Ci、P、Fr、Fo、G 分别代表 City、Province、Friends、Followers、Gender 属性的值域见表 1。当 B={Ci,G}时，则有:

$$U/SIM(B) = \{S_B(u_1) = \{1,3\}, S_B(u_2) = \{2,3,6\}, S_B(u_3) = \{1,2,3,6\}, S_B(u_4) = \{4,5\}, S_B(u_5) = \{4,5,6\}, S_B(u_6) = \{2,3,5,6\}\}$$

$$C_{\text{block}}(B) = \{X_1 = \{1,3\}, X_2 = \{2,3,6\}, X_3 = \{4,5\}, X_4 = \{5,6\}\}$$

表 1 不完备谣言信息表

Tabel 1 Incomplete rumor information table

ID	City	Province	Friends	Followers	Gender	d
1	005	014	[10,100]	[100,200]	Women	1
2	003	*	[10,100]	[100,200]	Women	1
3	*	*	[10,100]	[10,99]	Women	0
4	005	*	[10,100]	[10,99]	Men	1
5	*	*	[0,9]	[10,99]	Men	0
6	003	006	[10,100]	[100,200]	*	1

4 基于粗糙集的知识获取与决策

本文首先总结了粗糙集中用于不确定性度量的方式，以及它们在完备和不完备信息系统中的一致性表达，以此说明在完备信息系统中构造的知识约简与决策算法与在不完备信息系统中是一致的，在不完备信息系统中进行知识约简与决策更具有普适性和泛化性。

在此基础上，总结了 9 种用于构造约简算法的重要性度量，并厘清了它们之间存在的约束关系，实质上只能构造 4 种约简算法。本文从熵度量的角度提出了最大相关最小冗余约简算法，用于解决谣言信息系统的冗余性；并基于极大相容块提出了用决策粗糙指标构造的决策树算法，用于不完备谣言信息系统中的决策。框架如图 1 所示。

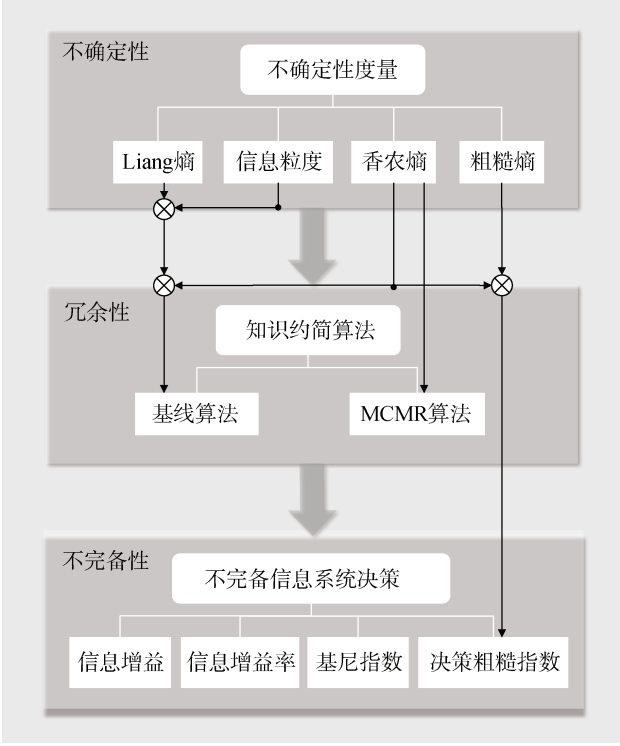


图 1 知识获取与决策框架

Figure 1 Knowledge acquisition and decision-making framework

4.1 不确定性度量

本文主要从不完备谣言信息系统的知识粗糙性方面刻画数据的不确定性，知识是利用属性分类的能力，随着属性的增多，对于论域中对象的分类就会越加精细。如例 1，利用属性 City 对论域进行划分，只能划分为 {1,3,4,5} 和 {2,3,5,6}；利用属性集 B = {Ci,G} 对论域进行划分，则能划分为 $\{1,3\}$, $\{2,3,6\}$, $\{4,5\}$, $\{5,6\}$。随着属性的增多，论域中的对象划分的越来越细，知识的粗糙性也越来越低。

在粗糙集理论中，度量不确定性有两个主要方向：粒度度量、熵度量。它们为从不同角度研究不确定性提供了相应的指标[22, 53]。粒度度量指标主要有 Liang 熵及信息粒度，以代数论集合观点定义；熵度

量指标主要有香农熵和粗糙熵, 从信息论信息观点定义。本文总结了这四种度量指标及其变体在完备和不完备信息系统中的一致性表达。具体见表 2, 相关符号说明见下述证明。

Liang 等人[22]已经证明了香农熵、Liang 熵、粗糙熵、信息粒度从完备信息系统到不完备信息系统的一致性拓展。本文推导证明了 Liang 熵中的条件信息熵在不完备信息系统中的一致性拓展, 其他变体如互信息熵、联合熵、互信息粒度、联合信息粒度等不确定度量方式在完备和不完备信息系统中的一致性证明类似, 这里不做一一证明。

定义 2. 设 $S=(U,A)$ 是一个不完备系统, $B,D\subseteq A$, $U/SIM(B)=\left\{S_B(u_1),S_B(u_2),...,S_B\left(u_{|U|}\right)\right\}$, $U/SIM(D)=\left\{S_D(u_1),S_D(u_2),...,S_D\left(u_{|U|}\right)\right\}$。

从 Liang 熵的角度, D 相对于 B 的条件信息熵定义为:

$$E(D/B)=\frac{1}{|U|}\sum_{i=1}^{|U|}\left(\frac{\left|S_B(u_i)-S_D(u_i)\right|}{|U|}\right) \tag{1}$$

引理 1[32]. 设 $S=(U,A)$ 是一个完备信息系统, $B,D\subseteq A$, $U/IND(B)=\{X_1,X_2,...,X_m\}$, $U/IND(D)=\{Y_1,Y_2,\cdots,Y_n\}$。 D 相对于 B 的条件信息熵退化为:

$$E(D/B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left|X_i\cap Y_j\right|}{|U|}\cdot\frac{\left|Y_j^c-X_i^c\right|}{|U|} \tag{2}$$

其中, X_i^c 和 Y_j^c 分别是 X_i 和 Y_j 的补集。

证明: 显然有 $X_i\cap Y_j\subseteq X_i$, $\bigcup_{j=1}^{n}(X_i\cap Y_j)=X_i$, $\bigcap_{j=1}^{n}(X_i\cap Y_j)=\varnothing$。 $X_i\cap Y_j$ 中的元素对应于 X_i 中的 $\left\{u_{ij}^1,u_{ij}^2,...,u_{ij}^{z_{ij}}\right\}$, $i=1,2,...,m$, $j=1,2,...,n$。 z_{ij} 代表集合 $X_i\cap Y_j$ 中的元素数量。 $X_i=\left\{u_i^1,u_i^2,...,u_i^{b_i}\right\}$, b_i 代表集合 X_i 中的元素数量, 并且有 $\sum_{j=1}^{n}z_{ij}=b_i$, $\sum_{i=1}^{m}b_i=|U|$。

因此,

$$X_i=S_B\left(u_{ij}^1\right)=S_B\left(u_{ij}^2\right)=...=S_B\left(u_{ij}^{z_{ij}}\right)$$

$$Y_j=S_D\left(u_{ij}^1\right)=S_D\left(u_{ij}^2\right)=...=S_D\left(u_{ij}^{z_{ij}}\right)$$

$$\frac{|X_1\cap Y_1|}{|U|}\cdot\frac{\left|Y_1^c-X_1^c\right|}{|U|}==\frac{|X_1\cap Y_1|}{|U|}\cdot\frac{|X_1-Y_1|}{|U|}$$
$$=\frac{1}{|U|}\frac{\left|S_B\left(u_{11}^1\right)-S_D\left(u_{11}^1\right)\right|}{|U|}+\frac{1}{|U|}\frac{\left|S_B\left(u_{11}^2\right)-S_D\left(u_{11}^2\right)\right|}{|U|}$$
$$+...+\frac{1}{|U|}\frac{\left|S_B\left(u_{11}^{z_{11}}\right)-S_D\left(u_{11}^{z_{11}}\right)\right|}{|U|}$$

那么,

$$\sum_{j=1}^{n}\frac{|X_1\cap Y_j|}{|U|}\cdot\frac{\left|Y_j^c-X_1^c\right|}{|U|}=\frac{1}{|U|}\frac{\left|S_B\left(u_{11}^1\right)-S_D\left(u_{11}^1\right)\right|}{|U|}$$
$$+\frac{1}{|U|}\frac{\left|S_B\left(u_{11}^2\right)-S_D\left(u_{11}^2\right)\right|}{|U|}+...+\frac{1}{|U|}\frac{\left|S_B\left(u_{11}^{z_{11}}\right)-S_D\left(u_{11}^{z_{11}}\right)\right|}{|U|}$$
$$+...+\frac{1}{|U|}\frac{\left|S_B\left(u_{12}^{z_{12}}\right)-S_D\left(u_{12}^{z_{12}}\right)\right|}{|U|}+...+\frac{1}{|U|}\frac{\left|S_B\left(u_{1n}^{z_{1n}}\right)-S_D\left(u_{1n}^{z_{1n}}\right)\right|}{|U|}$$
$$=\frac{1}{|U|}\frac{\left|S_B\left(u_1^1\right)-S_D\left(u_1^1\right)\right|}{|U|}+\frac{1}{|U|}\frac{\left|S_B\left(u_1^2\right)-S_D\left(u_1^2\right)\right|}{|U|}$$
$$+...+\frac{1}{|U|}\frac{\left|S_B\left(u_1^{b_1}\right)-S_D\left(u_1^{b_1}\right)\right|}{|U|}$$

所以,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|X_i\cap Y_j|}{|U|}\cdot\frac{\left|Y_j^c-X_i^c\right|}{|U|}=\frac{1}{|U|}\frac{\left|S_B\left(u_1^1\right)-S_D\left(u_1^1\right)\right|}{|U|}$$
$$+\frac{1}{|U|}\frac{\left|S_B\left(u_1^2\right)-S_D\left(u_1^2\right)\right|}{|U|}+...+\frac{1}{|U|}\frac{\left|S_B\left(u_1^{b_1}\right)-S_D\left(u_1^{b_1}\right)\right|}{|U|}$$
$$+...+\frac{1}{|U|}\frac{\left|S_B\left(u_2^{b_2}\right)-S_D\left(u_2^{b_2}\right)\right|}{|U|}+\frac{1}{|U|}\frac{\left|S_B\left(u_m^{b_m}\right)-S_D\left(u_m^{b_m}\right)\right|}{|U|}$$
$$=\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{\left|S_B(u_i)-S_D(u_i)\right|}{|U|}$$

引理(1)说明完备信息系统的条件信息熵是不完备信息系统的一个特殊实例。或者说不完备信息系统的条件信息熵是对完备信息系统的一致性扩展。

4.2 知识约简算法

4.2.1 属性约简及属性重要性

知识约简, 即属性约简[54], 是粗糙集理论中的热点研究, 旨在保持分类能力不变或者降低不多的情况下, 删除其中冗余或者不重要的属性。决策信息系统经过属性约简, 可以降低特征属性的维度[55], 提高基于属性的分类性能[56], 简化数据描述[57]和避免过拟合问题[58]。信息系统中属性约简的关键是如何构造系统中属性重要性的评价标准。

给定决策信息系统且 $B\subseteq C$，一般的，属性 $a\in B$ 的相对重要性定义为:

$$Sig(a,B,D)=M(B-\{a\},D)-M(B,D)$$

本文给出了新的定义:

$$Sig(a,B,D)=M(C,D)-M(B+\{a\},D) \tag{3}$$

$M(B,D)$ 为 B 相对于 D 的重要性度量。这两种定义在筛选约简子集的核属性时效果一致，但是前一公式是从属性自身相对于候选属性子集的重要性考虑，取其中的最大值；后一公式是相对于全属性集的度量距离来衡量，取其中的最小值，在算法实现上更容易控制阈值来选择核属性的个数。

当前，重要性度量主要从粒度度量和熵度量两个方面。从表 2 可以知道，$M(B,D)$ 有 9 种基本度量方式，分别是:

CH, MH, JH, CE, ME, JE, CG, MG, JG

其中 CH, MH, JH 是从熵度量角度, CE, ME, JE, CG, MG, JG 是从粒度度量角度。

跟决策有关的度量方式满足的约束条件有以下几种情况:

(1) 信息粒度和 Liang 熵的变体满足:

$$\begin{aligned}&E(D/B)=CG(B/D)\\&E(D;B)+MG(D;B)=1\\&E(D\cup B)+JG(D\cup B)=1\end{aligned} \tag{4}$$

表 2　不同度量在完备和不完备信息系统中的表达

Tabel 2　Expression of different measures in both complete and incomplete information systems

度量类	度量方法	完备信息系统	不完备信息系统
香农熵	信息熵	$H(B)=-\sum_{i=1}^{m}\frac{\lvert X_i\rvert}{\lvert U\rvert}\log_2\frac{\lvert X_i\rvert}{\lvert U\rvert}$	$H(B)=-\sum_{i=1}^{\lvert U\rvert}\frac{1}{\lvert U\rvert}\log_2\frac{\lvert S_B(u_i)\rvert}{\lvert U\rvert}$
	条件信息熵	$CH(D/B)=-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\log_2\frac{\lvert X_i\cap Y_j\rvert}{\lvert X_i\rvert}$	$CH(D/B)=-\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\log_2\frac{\lvert S_D(u_i)\cap S_B(u_i)\rvert}{\lvert S_B(u_i)\rvert}$
	互信息熵	$MH(D;B)=-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\log_2\frac{\lvert X_i\rvert\cdot\lvert Y_j\rvert}{\lvert X_i\cap Y_j\rvert\cdot\lvert U\rvert}$	$MH(D;B)=-\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\log_2\frac{\lvert S_D(u_i)\rvert\cdot\lvert S_B(u_i)\rvert}{\lvert S_D(u_i)\cap S_B(u_i)\rvert\cdot\lvert U\rvert}$
	联合熵	$JH(D\cup B)=-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\log_2\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}$	$JH(D\cup B)=-\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\log_2\frac{\lvert S_D(u_i)\cap S_B(u_i)\rvert}{\lvert U\rvert}$
Liang 熵	信息熵	$E(B)=\sum_{i=1}^{m}\frac{\lvert X_i\rvert}{\lvert U\rvert}\left(1-\frac{\lvert X_i\rvert}{\lvert U\rvert}\right)$	$E(B)=\sum_{i=1}^{\lvert U\rvert}\frac{1}{\lvert U\rvert}\left(1-\frac{\lvert S_P(u_i)\rvert}{\lvert U\rvert}\right)$
	条件信息熵	$CE(D/B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\cdot\frac{\lvert Y_j^c-X_i^c\rvert}{\lvert U\rvert}$	$CE(D/B)=\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\left(\frac{\lvert S_B(u_i)-S_D(u_i)\rvert}{\lvert U\rvert}\right)$
	互信息熵	$ME(D;B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\cdot\frac{\lvert Y_j^c\cap X_i^c\rvert}{\lvert U\rvert}$	$ME(D;B)=\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\left(\frac{\lvert(U-S_B(u_i))\cap(U-S_D(u_i))\rvert}{\lvert U\rvert}\right)$
	联合熵	$JE(D\cup B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\cdot\left(1-\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\right)$	$JE(D\cup B)=\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\left(1-\frac{\lvert S_B(u_i)\cap S_D(u_i)\rvert}{\lvert U\rvert}\right)$
粗糙熵	粗糙熵	$E_r(B)=-\sum_{i=1}^{m}\frac{\lvert X_i\rvert}{\lvert U\rvert}\log_2\frac{1}{\lvert X_i\rvert}$	$E_r(B)=-\sum_{i=1}^{\lvert U\rvert}\frac{1}{\lvert U\rvert}\log_2\frac{1}{\lvert S_B(u_i)\rvert}$
信息粒度	知识粒度	$G(B)=\frac{1}{\lvert U\rvert^2}\sum_{i=1}^{m}\lvert X_i\rvert^2$	$G(B)=\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\frac{\lvert S_B(u_i)\rvert}{\lvert U\rvert}$
	条件信息粒度	$CG(D/B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\cdot\frac{\lvert X_i^c-Y_j^c\rvert}{\lvert U\rvert}\right)$	$CG(D/B)=\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\left(\frac{\lvert S_D(u_i)\rvert-\lvert S_B(u_i)\cap S_D(u_i)\rvert}{\lvert U\rvert}\right)$
	互信息粒度	$MG(D;B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\lvert X_i\cap Y_j\rvert}{\lvert U\rvert}\cdot\frac{\lvert X_i\cup Y_j\rvert}{\lvert U\rvert}\right)$	$MG(D;B)=\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\left(\frac{\lvert S_D(u_i)\rvert+\lvert S_B(u_i)\rvert-\lvert S_B(u_i)\cap S_D(u_i)\rvert}{\lvert U\rvert}\right)$
	联合信息粒度	$JG(D\cup B)=\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\lvert X_i\cap Y_j\rvert^2}{\lvert U\rvert^2}\right)$	$JG(D\cup B)=\frac{1}{\lvert U\rvert}\sum_{i=1}^{\lvert U\rvert}\left(\frac{\lvert S_B(u_i)\cap S_D(u_i)\rvert}{\lvert U\rvert}\right)$

(2) 香农熵及其变体满足:

$$H(D;B)=H(D)-H(D/B) \tag{5}$$

(3) 信息粒度及其变体满足:

$$JG(D\cup B)=G(D)-CG(D/B) \tag{6}$$

以属性重要性选取相应属性加入约简属性集合时, 我们有以下推论:

- 由式 4 可知, Liang 熵变体与信息粒度变体是一致的, 即 CE 与 CG, ME 与 MG, JE 与 JG 是一致的;
- 由式 5 可知, 香农互信息熵与条件信息熵是一致的, 即 MH 与 CH 是一致的;
- 由式 6 可知, 条件信息粒度与联合信息粒度是一致的, 即 JG 与 CG (本文算法使用的 CG 指 $CG(B/D)$)是一致的。

最终我们得到的 9 种重要性度量实质上可归纳为 MH, JH, MG, JG 这 4 种。

4.2.2 最大相关最小冗余度量

上节提到了 MH, JH, MG, JG 这 4 种重要性度量方式, 前两种从代数集合观出发, 考虑的是某属性对论域中确定分类子集的影响; 后两种从信息观出发, 考虑的是某属性对于论域中不确定分类子集的影响, 它比代数表示更加直观, 能够导出高效的知识约简算法[59]。本文从熵度量角度提出了最大相关最小冗余度量指标。

在图2中, D 表示决策属性集, B 表示约简属性集, a_i 表示候选属性, 可以得到:

$$①H(D;B/a_i)=MH(D;B\cup\{a_i\})-MH(D;a_i)$$

$$②H(D;B;a_i)=MH(D;B)-①H(D;B/a_i)$$

$$③\ H(D;a_i/B)=MH(D;B\cup\{a_i\})-MH(D;B)$$

在属性选择上, $MH(D;B\cup a_i)=①+②+③\triangleq ③$(这种情况下 B 确定)。在图 2 中, ③ 代表新增属性 a_i 对于决策贡献的信息量, 而 ② 代表了 a_i 与约简属性集 B 的重叠信息, 即对于决策冗余的信息量, 有时还会成为噪音数据。我们希望选到 ③ 尽可能大 ② 尽可能小的属性, 因此我们提出了最大相关最小冗余重要性度量指标 MCMR:

$$Sig(a_i,B,D)=\frac{\alpha\cdot ③-\beta\cdot ②}{H(a_i)} \tag{7}$$

这里我们称 ③ 为候选属性 a_i 的最大相关指标 MaxC, ② 为候选属性 a_i 的最小冗余指标 MinR, α,β 代表相应的权重值, 且 $\alpha+\beta=1$。$H(a_i)$ 为 a_i 的信息熵, 主要是为了避免倾向于选取取值较多的属性。

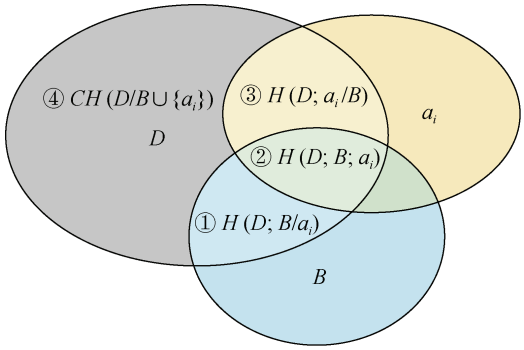


图 2 熵度量维恩图

Figure 2 Entropy measure Venn diagram

4.2.3 基于属性重要性的约简算法

给定不完备谣言决策信息系统 $(U,C\cup D)$ 且 $B\subseteq C$, 基于不同的属性重要性度量指标, 构造不同的约简算法。

基于 MH, JH, MG, JG 的知识约简算法, 如算法 1, 其中重要性度量指标 $M(B,D)$ 分别使用 MH, JH, MG, JG 进行度量。

算法 1.基于 MH, JH, MG, JG 的约简算法

输入: 不完备谣言信息系统 $(U,C\cup D)$。

输出: 一个属性约简子集 B

1: 计算重要性度量 $M(C,D)$;

2: 置初始属性约简子集 $B=\varnothing$。

3: WHILE $\left|M(C,D)-M(B,D)\right|>q$ DO

4: $\overline{a}_i=\min\limits_{a_i=C-B}\left|M(C,D)-M(B+\{a_i\},D)\right|$

5: $B=B\cup\{\overline{a}_i\}$

6: END WHILE

基于最大相关 MaxC ③ 、最小冗余 MinR ② 以及最大相关最小冗余指标 MCMR 的属性约简算法, 如算法 2 所示, 其中属性重要性 $Sig(a,B,D)$ 分别使用 $③H(D;a_i/B)$, $-②H(D;B;a_i)$ 以及 $\frac{\alpha\cdot ③-\beta\cdot ②}{H(a_i)}$ 进行度量。

其中算法 1 以及算法 2 中基于 MaxC、MinR 的约简算法代表了 6 类知识约简算法, 前 4 种是前文讲到的 4 种不同度量方式构造的算法, 是常用的约简算法; 后 2 种是基于不确定度量方式总结的 2 类算法, 分别代表了两种方向或者思考, 一种基于相关性, 一种基于冗余性; 而基于 MCMR 的约简算法是本文提出的算法。

算法 2.基于 MaxC, MinR, MCMR 的约简算法

输入: 不完备谣言信息系统$(U, C\cup D)$。

输出: 一个属性约简子集 B

1: 计算互信息 $MH(C,D)$;

2: 置初始属性约简子集 $B=\varnothing$。

3: WHILE $\left|MH(C;D)-MH(B;D)\right|>\theta$ Do

4: $\overline{a}_i=\max\limits_{a_i=C-B} Sig(a_i,B,D)$

5: $B=B\cup\{\overline{a}_i\}$

6: END WHILE

如例 1, 假设这里以算法 1 中的基于 MH 构建的约简算法来选择约简属性集。以属性 Fo 划分的等价和相容关系类分别为:

$$U/IND(Fo)=\{X_1=\{1,2,6\},X_2=\{2,3,5\}\}$$

$$\begin{aligned}U/SIM(Fo)=\{&S_{Fo}(u_1)=\{1,2,6\},\\&S_{Fo}(u_2)=\{1,2,6\},\\&S_{Fo}(u_3)=\{3,4,5\},\\&S_{Fo}(u_4)=\{3,4,5\},\\&S_{Fo}(u_5)=\{3,4,5\},\\&S_{Fo}(u_6)=\{1,2,6\}\}\end{aligned}$$

以决策属性 d 划分的等价和相容关系类分别为:

$$U/IND(d)=\{Y_1=\{1,2,4,6\},Y_2=\{3,5\}\}$$

$$\begin{aligned}U/SIM(d)=\{&S_d(u_1)=\{1,2,4,6\},\\&S_d(u_2)=\{1,2,4,6\},\\&S_d(u_3)=\{3,5\},\\&S_d(u_4)=\{1,2,4,6\},\\&S_d(u_5)=\{3,5\},\\&S_d(u_6)=\{1,2,4,6\}\}\end{aligned}$$

则 $\{Fo\cup d\}$ 划分的等价和相容关系类为:

$$U/IND(Fo\cup d)=\{\{1,2,6\},\{4\},\{3,5\}\}$$

$$\begin{aligned}U/SIM(Fo\cup d)=\{&S_{Fo\cup d}(u_1)=\{1,2,6\},\\&S_{Fo\cup d}(u_2)=\{1,2,6\},\\&S_{Fo\cup d}(u_3)=\{3,5\},S_{Fo\cup d}(u_4)=\{4\},\\&S_{Fo\cup d}(u_5)=\{3,5\},\\&S_{Fo\cup d}(u_6)=\{1,2,6\}\}\end{aligned}$$

计算得到互信息熵为:

$$\begin{aligned}MH(d;Fo)&=-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left|X_i\cap Y_j\right|}{|U|}\log_2\frac{|X_i|\cdot|Y_j|}{\left|X_i\cap Y_j\right|\cdot|U|}\\&=\frac{1}{2}\log_2 3-\frac{1}{3}\approx 0.459\end{aligned}$$

$$\begin{aligned}MH(d;Fo)&=-\frac{1}{|U|}\sum_{i=1}^{|U|}\log_2\frac{\left|S_d(u_i)\right|\cdot\left|S_{B_1}(u_i)\right|}{\left|S_d(u_i)\cap S_{B_1}(u_i)\right|\cdot|U|}\\&=\frac{1}{2}\log_2 3-\frac{1}{3}\approx 0.459\end{aligned}$$

可以看到, 两者结果相同, 也间接证明了 4.1 节中的香农互信息熵在不完备信息系统中的表达是完备信息系统的一致性拓展。同理分别计算出以属性 Ci, P, Fr, G 划分论域时的互信息熵, 选取其中互信息熵最大的作为核属性加入到约简子 B 中。并按照算法 1 得到最终的属性约简子集 B。

4.3 不完备决策树算法

对于一个非完备的信息系统来说, 由于完备系统与非完备系统在拓扑结构上存在本质差异[51, 60-61], 所以在非完备系统中, 香农熵定义所需要的离散型概率分布也是不可能得到的, 只能通过相容关系或者极大相容块进行拓展。如例 1 中, 在属性 Ci, P, G 等列的属性值存在缺失, 只能采取传统的数值填补或者删除, 不能直接用传统意义的完备信息系统的方法来计算各类熵值。因此, 本文基于极大相容块提出决策粗糙指数, 并构建了决策树模型用于不完备系统中的谣言检测。

定义 3[62]. 设 U 是一个论域, B 是一个条件属性集合, $B\subseteq C$, $U/B=\{X_1,X_2,...,X_m\}$, D 为决策属性, 其中 $U/B=\{D_1,D_2,...,D_t\}$ 为决策概念集, 则决策概念集的粗糙熵为

$$E(D_B)=-\sum_{i=1}^{m}\sum_{j=1}^{t}\frac{|X_i|}{|U|}\log\frac{\left|X_i\cap D_j\right|}{|X_i|}\tag{8}$$

由 2.2 节可知 $SIM(P)=\bigcap\limits_{a\in P}SIM(\{a\})$, 并且在不完备信息系统中, 相似关系也是相容关系。基于极大相容块, 将定义 3 扩展到不完备信息系统。

算法 3.不完备决策树算法

输入: 不完备谣言信息系统$(U,C\cup D,f)$, 阈值 $\varepsilon>0$

输出: 一个决策树 T

1: 计算属性约简集合 B; ▷ MCMR 算法 2

2: $(U,C\cup D,f)\Leftarrow(U,B\cup D,f)$

3: FOR $a_i\in C$ DO

$a=\underset{a_i\in C}{arg\min}\left(D(a_i)\right)$ ▷ 决策粗糙指数

$C\Leftarrow C-\{a\}$

4: IF $D(a) \leqslant \varepsilon$ DO

5: $C_{\text{block}}(a)=\{X_1, X_2, ..., X_m\}$

▷ 极大相容块类

6: FOR $X_j \in C_{\text{block}}(a)$ DO

$(U, C\cup D, f) \Leftarrow (X_j, C\cup D, f)$

7: $C_{\text{block}}(D)=\{D_1, D_2, ..., D_j, ..., D_n\}$

8: IF $|C_{\text{block}}(D)|$=1 THEN

$Label_T \Leftarrow f(U, D)$

9: ELSE IF $C=\varnothing$ THEN

$Label_T \Leftarrow \underset{f(U,D_j)}{arg\ \max} |D_j|$

10: ELSE

RETURN Step3

END IF

END FOR

END IF

END FOR

11: RETURN T

$$C_{\text{block}}(B)=\{X_1, X_2, ..., X_m\},$$
$$C_{\text{block}}(D)=\{D_1, D_2, ..., D_t\},$$
$$C_{\text{block}}(B\cup D)=\{Y_1, Y_2, ..., Y_n\}$$

其中 Y_j 为属性 $B\cup D$ 所决定的极大相容块。

则有:

$$E(D_B)=-\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{|X_i|}{|U|}\log\frac{|Y_j|}{|X_i|} \tag{9}$$

属性 B 在 U 上导出的极大相容类 $C_{block}(B)$ 的香农信息熵为

$$H(B)=-\sum_{i=1}^{m}\frac{|X_i|}{|U|}\log_2\frac{|X_i|}{|U|} \tag{10}$$

决策粗糙熵更倾向于区分属性的分类能力, 决策粗糙熵越小, 分类粒度越细, 但分类效果并不一定好。而条件熵 $H(D/B)$ 能够衡量决策属性在条件属性集 B 下的信息熵, 条件熵越小, 说明条件属性集 B 消除的不确定性越大。结合决策粗糙熵以及条件熵的优势, 我们提出了决策粗糙指数 $D(B)$。

$$D(B)=\frac{\alpha\cdot E(D_B)+\beta\cdot H(D/B)}{H(B)} \tag{11}$$

α, β 代表相应的权重值, 且 $\alpha+\beta=1$。决策粗糙指数 $D(B)$ 越小, 选用属性集 B 用于决策的分类效果越好。

基于决策粗糙指数 $D(B)$ 的不完备决策树算法, 如算法 3 所示。

5 实验结果分析

本文对 4.2.3 节和 4.3 节所提到的算法进行了实验论证。

5.1 相关数据集

本文使用了 4 个 UCI 机器学习数据集用于 4.2.3 节中 MCMR 指标与其他重要性度量的对比实验, 这 4 个数据集分别是 Car、Abalone、Adult 以及 Bank, 具体实验在 5.2 节。本文从 Weibo 数据集[63-64]中提取了 Post、Structure、User、Union 等 4 个数据集。在 5.3 节中检验了本文所提约简算法在不同模型上的分类效果以及约简效果, 这些模型有最大熵、决策树、提升算法+决策树以及随机森林等, 它们的核心都使用了不同的不确定性度量方式。在 5.4 节中对比了本文所提算法 3 与传统决策树在不完备信息系统中的决策分类效果。数据集描述见表 3。

表 3 数据集描述

Tabel 3 Description of the dataset

数据集	属性数	对象数	类别数
Car	6	1728	2
Abalone	8	4177	2
Adult	14	3256	2
Bank	16	4521	2
Post	8	4664	2
Structure	5	4664	2
User	12	4664	2
Union	25	4664	2

特别地, Post 数据集为 Weibo 贴文元数据以及本文基于知网情感词典从文本中提取的文本长度和情感词, 包括积极词个数、积极评分, 消极词个数以及消极评分。Structure 数据集是本文通过 Neo4j 图数据库提取的贴文在传播过程中的简单结构特征, 包括传播最大半径, 最大尺寸所在层, 最大尺寸, 节点数, 边数。User 数据集是 Weibo 数据集中自带的统计特征, Union 数据集是上述 3 种数据集的联合。关于 Weibo Union 数据集中各个属性的名称说明如表格 4 所示。

5.2 知识约简算法对比分析

为了更好地检验约简算法的有效性以及降低时间复杂度, 本文随机抽取 10%的数据用于约简, 剩余

表 4　Weibo 数据集属性描述

Tabel 4　Description of the attributes in Weibo datasets

属性名	含义	数据集
bi_followers_count	双向关注数	User
attitudes_count	点赞数	Post
city	所在城市	User
favourites_count	收藏数	User
comments_count	评论数	Post
followers_count	粉丝数	User
friends_count	关注数	User
Max_radius	最大半径	Structure
gender	性别	User
Negnum	消极词个数	Post
Node_num	节点个数	Structure
Max_size	最大尺寸	Structure
Len	文本长度	Post
Negscore	消极评分	Post
Max_size_layer	最大尺寸层	Structure
province	省份	User
Posscore	积极评分	Post
Posnum	积极词个数	Post
Rel_num	关系(边)数	Structure
reposts_count	转发数	Post
statuses_count	微博数	User
user_created	发文时间跨度	User
user_geo_enabled	位置是否可用	User
user_location	用户定位位置	User
verified_type	账户认证类型	User

数据用于检验约简效果。本节所使用的分类模型为决策树模型，分类时采用五折交叉验证的方法。

本节在 Car,Abalone,Adult 以及 Bank 等 4 个数据集上对比了 MCMR 指标与其他重要性度量。结果如图 3 所示，不同曲线表示采用不同重要性度量约简到统一属性数量时能达到的分类效果，同时图例中列举了所有的重要性度量指标，并把效果一致的指标划归到一组。其中 NoAttriRedu 表示不使用约简算法即使用全属性进行分类时的效果。

如图 3，除了在 Car 数据集上不使用约简算法时谣言识别的效果更好，在其他几个数据集上，约简算法不仅能够减小分类模型的计算开销，还能提升识别效果，这表明 Car 数据集的属性之间的冗余很少，每个属性对于决策分类都很重要。并且从算法对比上看，本文所提 MCMR 指标优于其他几种基线指标，能够更大地保留信息，去除冗余和噪音，提高分类效果。

5.3　约简前后谣言检测对比分析

本节将 MCMR 指标应用于谣言检测。在检测前，首先进行了 Union 数据集的分析。本文对两类标签下每种属性的分布进行了分析，如图 4 所示，其中红色分布为谣言的概率密度分布，蓝色分布为非谣言的概率密度分布。不同标签的概率密度分布在一些属性特征上相差极大，也就是说，这些属性特征用于分类的效果要比较好。

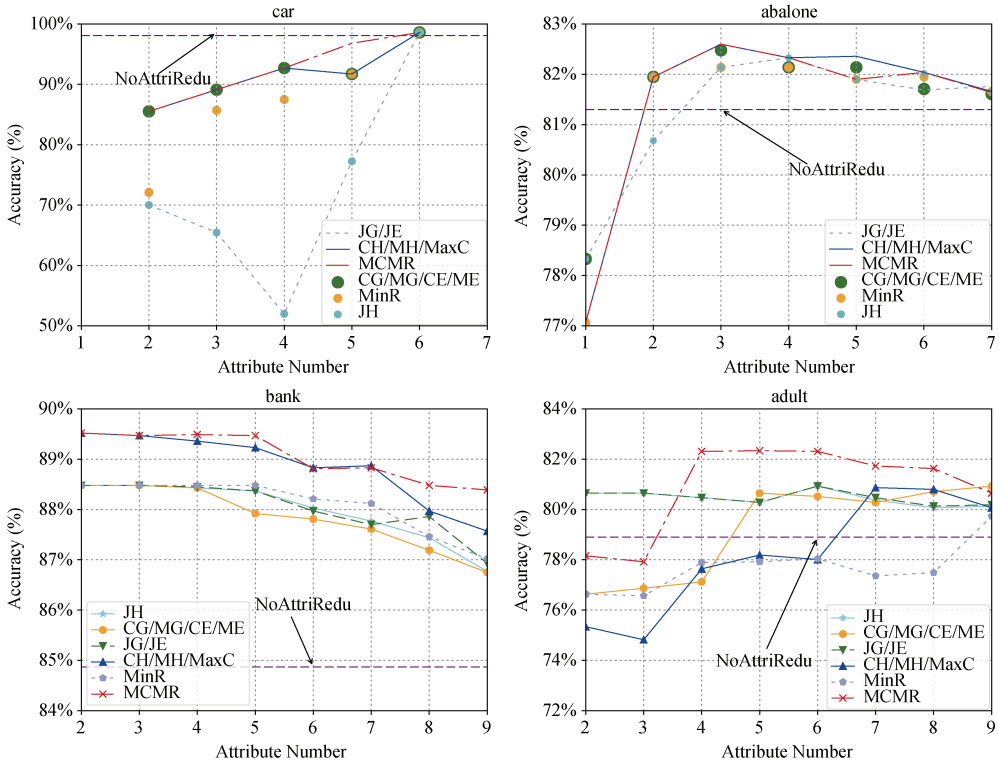


图 3　属性约简算法对比

Figure 3　Attribute reduction algorithms comparison

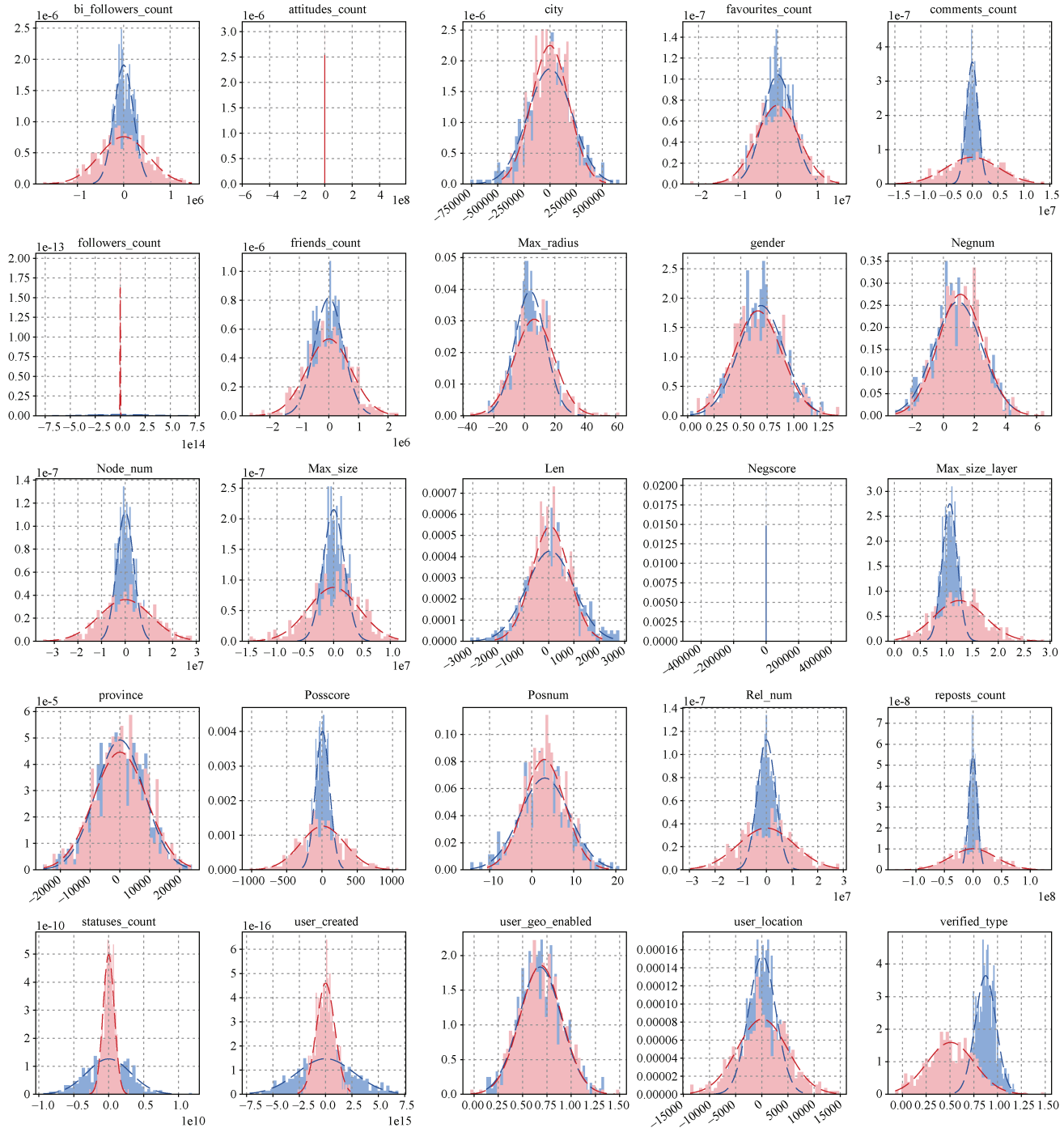


图 4　Union 数据集分析

Figure 4　Union dataset analysis

以 attitudes_count, Negscore 为例。attitudes_count 代表了点赞数，谣言的点赞数集中在 0 附近，而非谣言的点赞数分布更加均匀。Negscore 代表了文本的消极情绪评分，非谣言的消极情感评分集中在 0 附近，而谣言的消极情感评分分布更加均匀。两种标签的分布相差大的属性更适合用于谣言识别。

本文在几种基于不确定度量的分类模型上检验了约简前后谣言识别的效果，这几种模型有最大熵模型，决策树模型，AdaBoost+决策树模型，随机森林模型。与 5.2 节不同的是我们虽然同样随机选取了 10%数据用于属性约简，但是把整个数据集用于谣言分类。实验结果见表 5～表 8，前表示约简前，后表示约简后，红色表示约简后检测效果提升，绿色表示下降。

从表 5～表 7，可以看出以 MCMR 指标构建的约简算法不仅能够较好地约简属性，减小检测模型的计算开支，还能消除冗余与噪音，提高谣言检测的

效果。在 Union 数据集中虽然没能提高谣言识别的效果, 但是分类模型的计算开支减少了一半。图 5 是 Union 数据集采用随机森林(Gini)约简前后的属性热力图, 从图中可以看到约简后的属性冗余更小, 表明本文所提算法能够有效解决不完备谣言信息系统的冗余性。

表 5　Post 数据集

Tabel 5　Post dataset

模型	属性数		计算时间(s)		准确率		召回率		F1 值	
	前	后	前	后	前	后	前	后	前	后
最大熵(IIS)	8	3	14.70	8.07	0.7832	0.8132	0.8088	0.8400	0.7872	0.8170
最大熵(GIS)	8	3	11.89	7.90	0.7755	0.8049	0.8127	0.8426	0.7821	0.8108
ID3	8	7	0.16	0.12	0.8700	0.8713	0.8638	0.8651	0.8683	0.8696
Cart	8	6	0.13	0.08	0.8636	0.8681	0.8560	0.8629	0.8615	0.8664
AdaBoost + ID3	8	2	3.73	2.88	0.8996	0.9039	0.9438	0.9542	0.9031	0.9078
AdaBoost + Cart	8	2	3.48	2.58	0.8986	0.9035	0.9425	0.9529	0.9021	0.9073
RandomForest (entropy)	8	6	2.43	2.27	0.9080	0.9093	0.9446	0.9434	0.9105	0.9116
RandomForest (gini)	8	7	1.88	1.69	**0.9095**	**0.9108**	**0.9459**	**0.9459**	**0.9120**	**0.9131**

表 6　Structure 数据集

Tabel 6　Structure dataset

模型	属性数		计算时间(s)		准确率		召回率		F1 值	
	前	后	前	后	前	后	前	后	前	后
最大熵(IIS)	5	3	11.47	7.87	0.7159	0.7433	0.7526	0.7794	0.7242	0.7507
最大熵(GIS)	5	2	11.18	7.67	0.7056	0.7412	0.7357	0.7816	0.7124	0.7497
ID3	5	4	0.10	0.08	0.7572	0.7521	0.7565	0.7474	0.7556	0.7492
Cart	5	4	0.07	0.06	0.7450	0.7473	0.7366	0.7448	0.7412	0.7449
AdaBoost + ID3	5	4	3.24	2.83	0.7988	0.7991	0.7837	0.7842	0.7944	0.7946
AdaBoost + Cart	5	4	3.12	2.80	0.8012	0.8012	0.7898	0.7898	0.7976	0.7976
RandomForest (entropy)	5	4	2.89	2.52	0.8042	**0.8029**	**0.8179**	**0.8175**	0.8056	**0.8045**
RandomForest (gini)	5	4	1.91	1.74	**0.8049**	0.8024	0.8166	0.8162	**0.8059**	0.8039

表 7　User 数据集

Tabel 7　User dataset

模型	属性数		计算时间(s)		准确率		召回率		F1 值	
	前	后	前	后	前	后	前	后	前	后
最大熵(IIS)	12	10	24.82	20.46	0.8619	0.8610	0.8914	0.8906	0.8648	0.8640
最大熵(GIS)	12	10	14.47	12.96	0.8730	0.8681	0.8923	0.8914	0.8745	0.8702
ID3	12	9	0.19	0.12	0.8934	0.8906	0.8893	0.8910	0.8921	0.8898
Cart	12	6	0.17	0.09	0.8900	0.8900	0.8884	0.8867	0.8890	0.8887
AdaBoost + ID3	12	4	5.37	3.83	0.9116	0.9136	0.9403	0.9442	0.9135	0.9155
AdaBoost + Cart	12	8	4.32	3.70	0.9110	0.9116	0.9373	0.9429	0.9126	0.9137
RandomFor-est(entropy)	12	7	2.72	2.04	**0.9219**	**0.9200**	**0.9581**	**0.9550**	**0.9241**	**0.9221**
RandomForest(gini)	12	10	2.05	1.90	0.9202	0.9196	0.9555	0.9524	0.9223	0.9216

表 8 Union 数据集

Tabel 8 Union dataset

模型	属性数		计算时间(s)		准确率		召回率		F1 值	
	前	后	前	后	前	后	前	后	前	后
最大熵(IIS)	25	14	63.51	25.61	0.8936	0.8692	0.9343	0.9066	0.8971	0.8730
最大熵(GIS)	25	12	23.91	14.53	0.8878	0.8703	0.9100	0.9005	0.8895	0.8732
ID3	25	13	0.24	0.12	0.9119	0.9016	0.9157	0.8992	0.9115	0.9006
Cart	25	9	0.31	0.09	0.9106	0.9003	0.9061	0.8945	0.9095	0.8990
AdaBoost + ID3	25	11	8.43	4.32	0.9301	0.9282	0.9468	0.9520	0.9307	0.9293
AdaBoost + Cart	25	9	7.69	3.60	0.9318	0.9256	0.9477	0.9468	0.9324	0.9266
RandomForest(entropy)	25	10	3.84	2.57	0.9402	0.9301	**0.9680**	0.9598	0.9414	0.9317
RandomForest(gini)	25	9	3.02	1.73	**0.9406**	**0.9312**	0.9676	**0.9619**	**0.9418**	**0.9328**

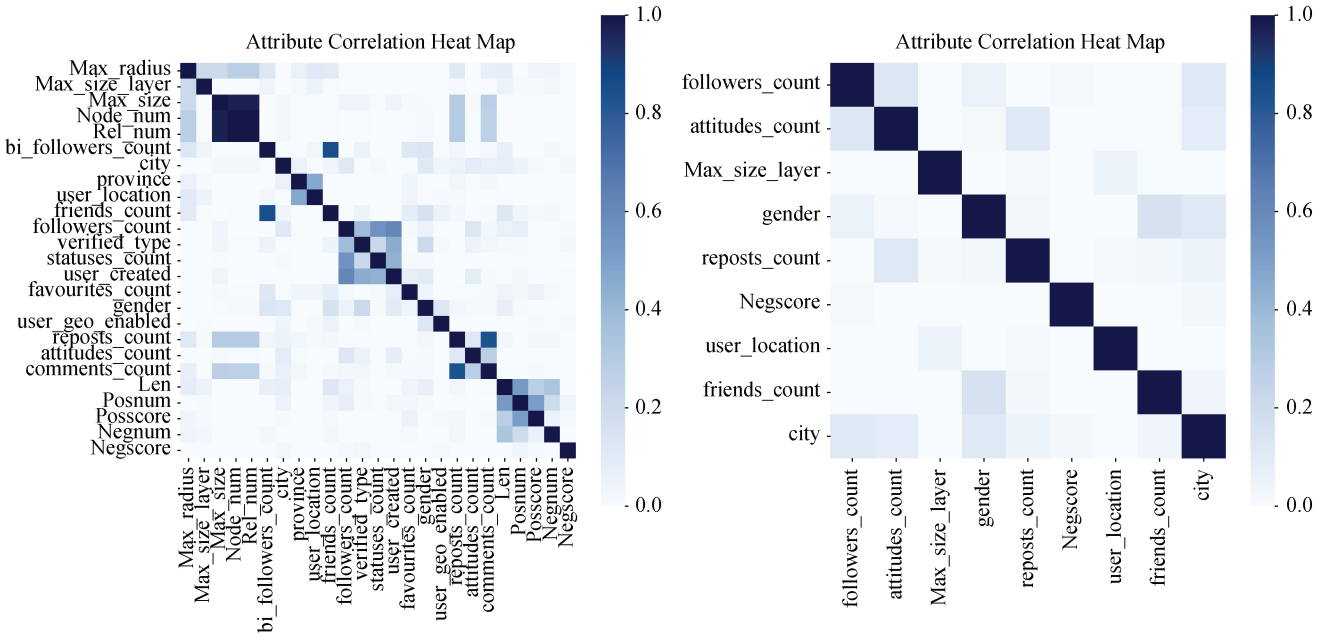


图 5 Union 数据集约简前后属性热力图

Figure 5 Union dataset heat map before and after attribute reductio

5.4 不完备信息中的决策对比分析

采用随机置空的方法, 将 Union 数据集的数据随机设置 20%～40%的空值。需要注意的是, 在随机置空时, 每个属性都会单独随机置空相应比例样本的属性值。传统的缺失值处理包括丢弃与填补。由于丢弃操作会损失一定的样本, 无法对丢失的样本进行谣言检测, 所以我们只与传统的填补处理进行了对比, 传统的填补处理包括填 0, 填 1 与填补样本平均值。我们对比了本文不完备决策树算法与传统的决策树算法在不完备谣言信息系统中识别谣言的效果, 如图 6, a～c 为本文不完备决策树算法与填补 0 的传统决策树算法的对比, d～f 为与填补 1 的决策树算法的对比, g～i 为与填补均值的决策树算法的对比, 横坐标代表了不同程度的数据缺失。

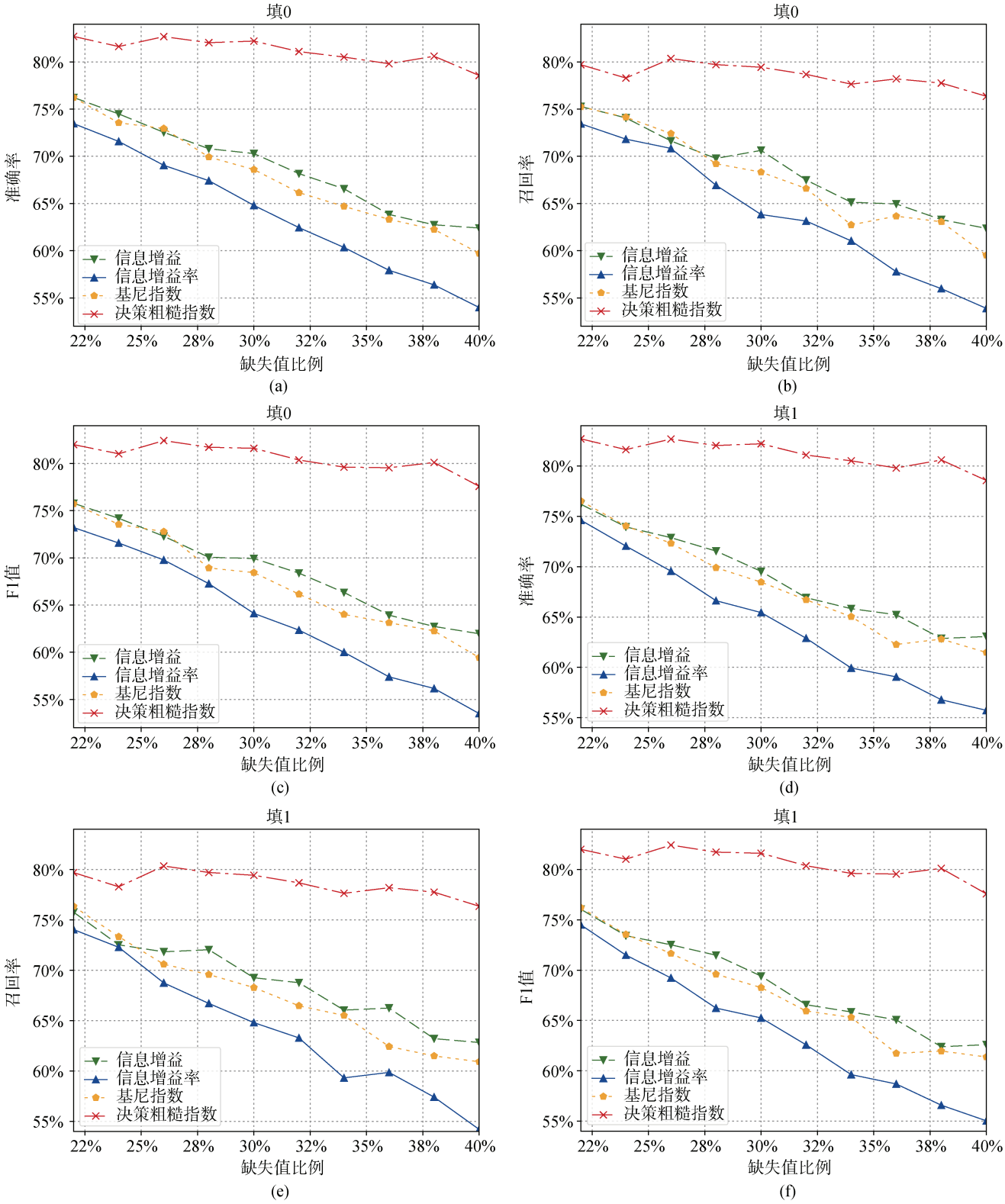
从图中可以看出, 采用基于极大相容块理论的不完备决策树算法在处理不完备数据时更加有效, 准确率、召回率、F1 值等明显高于传统填补操作。

6 结论

本文将粗糙集理论应用于不完备谣言信息系统进行知识获取与决策, 主要是通过解决不完备信息系统中的不确定性度量, 不完备性以及冗余性等问题来提高谣言数据的质量, 最终目的是为了提高谣言识别的效果。本文首先总结了粗糙集理论中用于不确定性度量的方法和它们在完备信息系统和不完备信息系统中的一致性表达。另外, 本文提出了一种

基于最大相关最小冗余的熵度量指标，并构造不完备信息系统中的属性约简算法。实验表明，本文算法优于几种基线算法，能够更好地在增加决策信息的同时减小冗余噪音，解决不完备谣言信息系统的冗余性问题。最后，本文基于极大相容块提出了一种以决策粗糙指数为核心的不完备决策树算法。实验表明，本文算法处理不完备信息的性能优于传统填补数据的决策树算法，能够有效地解决谣言信息系统的不完备性。

在下一步的研究中可以将知识约简以及不完备决策树模型以随机森林的形式进行集成，构造决策能力更强的分类模型。



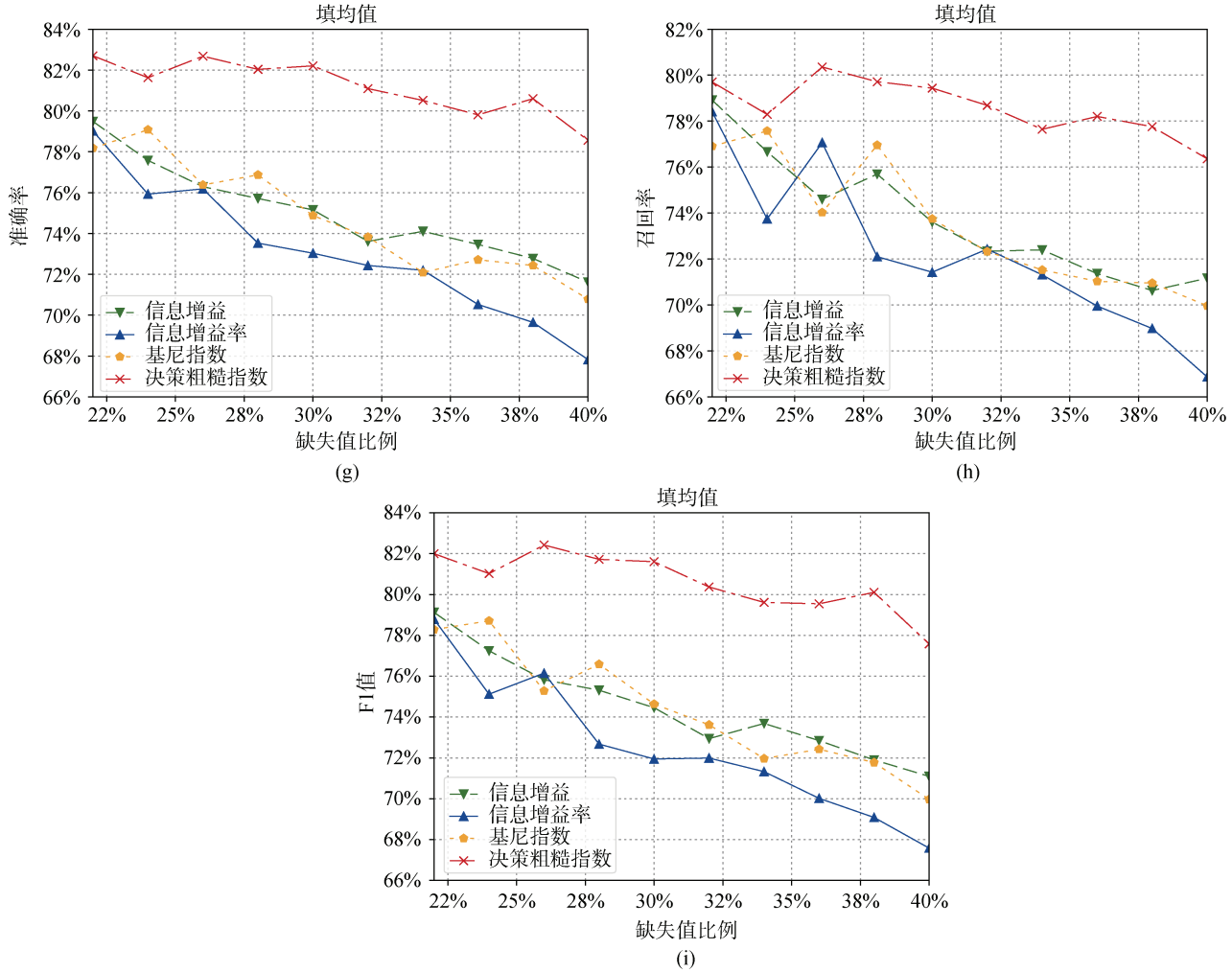


图 6　不完备谣言检测

Figure 6　Incomplete rumor detection

参考文献

[1] Yu S S, Li M C, Liu F M. Rumor Identification with Maximum Entropy in MicroNet[J]. *Complex*, 2017, 2017: 1703870: 1-1703870: 8.

[2] Liu S X, Ji X S, Liu C X, et al. A Complex Network Evolution Model for Network Growth Promoted by Information Transmission[J]. *Acta Physica Sinica*, 2014, 63(15): 429-439.
(刘树新, 季新生, 刘彩霞, 等. 一种信息传播促进网络增长的网络演化模型[J]. *物理学报*, 2014, 63(15): 429-439.)

[3] Shi H R, Ji L X, Liu S X, et al. Abnormal Link Detection Algorithm Based on Semi-Local Structure[J]. *Chinese Journal of Network and Information Security*, 2022, 8(1): 63-72.
(石灏苒, 吉立新, 刘树新, 等. 基于半局部结构的异常连边识别算法[J]. *网络与信息安全学报*, 2022, 8(1): 63-72.)

[4] Zhao Z, Resnick P, Mei Q Z. Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts[C]. *The 24th International Conference on World Wide Web*, 2015: 1395-1405.

[5] Kwon S, Cha M, Jung K, et al. Prominent Features of Rumor Propagation in Online Social Media[C]. *2013 IEEE 13th International Conference on Data Mining*, 2014: 1103-1108.

[6] Cai G Y, Wu H, Lv R. Rumors Detection in Chinese via Crowd Responses[C]. *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2014: 912-917.

[7] Yang F, Liu Y, Yu X H, et al. Automatic Detection of Rumor on Sina Weibo[C]. *The ACM SIGKDD Workshop on Mining Data Semantics*, 2012: 1-7.

[8] Wu K, Yang S, Zhu K Q. False Rumors Detection on Sina Weibo by Propagation Structures[C]. *2015 IEEE 31st International Conference on Data Engineering*, 2015: 651-662.

[9] Yang Z F, Wang C, Zhang F, et al. Emerging Rumor Identification for Social Media with Hot Topic Detection[C]. *2015 12th Web Information System and Application Conference*, 2016: 53-58.

[10] Takahashi T, Igata N. Rumor Detection on Twitter[C]. *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, 2013: 452-457.

[11] Chen Y X, Sui J, Hu L, et al. Attention-Residual Network with CNN for Rumor Detection[C]. *The 28th ACM International Conference on Information and Knowledge Management*, 2019: 1121-1130.

[12] Lv S, Zhang H, He H, et al. Microblog rumor detection based on comment sentiment and CNN-LSTM [M]. *Artificial Intelligence in China. Springer.* 2020: 148-156.

[13] Miao X, Rao D, Jiang Z. Syntax and Sentiment Enhanced BERT

for Earliest Rumor Detection[C]. *proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*, 2021: 570-582.

[14] Singh J P, Kumar A, Rana N P, et al. Attention-Based LSTM Network for Rumor Veracity Estimation of Tweets[J]. *Information Systems Frontiers*, 2022, 24(2): 459-474.

[15] Ma J, Gao W, Wong K F. Rumor Detection on Twitter with Tree-Structured Recursive Neural Networks[C]. *The 56th Annual Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), 2018: 1980-1989.

[16] Li L Z, Cai G Y, Chen N N. A Rumor Events Detection Method Based on Deep Bidirectional GRU Neural Network[C]. *2018 IEEE 3rd International Conference on Image*, *Vision and Computing*, 2018: 755-759.

[17] Wang Z H, Guo Y, Wang J H, et al. Rumor Events Detection from Chinese Microblogs via Sentiments Enhancement[J]. *IEEE Access*, 2019, 7: 103000-103018.

[18] Bian T A, Xiao X, Xu T Y, et al. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(1): 549-556.

[19] Lotfi S, Mirzarezaee M, Hosseinzadeh M, et al. Detection of Rumor Conversations in Twitter Using Graph Convolutional Networks[J]. *Applied Intelligence*, 2021, 51(7): 4774-4787.

[20] Jin Z W, Cao J, Guo H, et al. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs[C]. *The 25th ACM international conference on Multimedia*, 2017: 795-816.

[21] Salehi F, Keyvanpour M R, Sharifi A. SMKFC-ER: Semi-Supervised Multiple Kernel Fuzzy Clustering Based on Entropy and Relative Entropy[J]. *Information Sciences*, 2021, 547: 667-688.

[22] Liang J, Shi Z, Li D, et al. Information Entropy, Rough Entropy and Knowledge Granulation in Incomplete Information Systems[J]. *International Journal of General Systems*, 2006, 35(6): 641-654.

[23] Wang J, Wei J M, Yang Z L, et al. Feature Selection by Maximizing Independent Classification Information[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(4): 828-841.

[24] Zdzisaw Pawlak. Rough sets[J]. *International journal of parallel programming,* 1982, 11(5): 341-356.

[25] Wu Z Y, Pi D C, Chen J F, et al. Rumor Detection Based on Propagation Graph Neural Network with Attention Mechanism[J]. *Expert Systems With Applications*, 2020, 158: 113595.

[26] Song C G, Shu K, Wu B. Temporally Evolving Graph Neural Network for Fake News Detection[J]. *Information Processing & Management*, 2021, 58(6): 102712.

[27] He Z Y, Li C, Zhou F, et al. Rumor Detection on Social Media with Event Augmentations[C]. *The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021: 2020-2024.

[28] Düntsch I, Gediga G. Uncertainty Measures of Rough Set Prediction[J]. *Artificial Intelligence*, 1998, 106(1): 109-137.

[29] Beaubouef T, Petry F E, Arora G. Information-Theoretic Measures of Uncertainty for Rough Sets and Rough Relational Databases[J]. *Information Sciences*, 1998, 109(1/2/3/4): 185-195.

[30] Liang J Y, Xu Z B. Uncertainty Measures of Roughness of Knowledge and Rough Sets in Incomplete Information Systems[C]. *The 3rd World Congress on Intelligent Control and Automation* (*Cat. No.00EX393*), 2002: 2526-2529.

[31] Liang J, Xu Z, Miao D. Reduction of knowledge in incomplete information systems[C]. *proceedings of the Proceedings of Conference on Intelligent Information Processing in 16th World Computer Congress*, 2000, 7: 528-532.

[32] Liang J Y, Chin K S, Dang C Y, et al. A New Method for Measuring Uncertainty and Fuzziness in Rough Set Theory[J]. *International Journal of General Systems*, 2002, 31(4): 331-342.

[33] Liang J Y, Shi Z Z. The Information Entropy, Rough Entropy and Knowledge Granulation in Rough Set Theory[J]. *International Journal of Uncertainty*, *Fuzziness and Knowledge-Based Systems*, 2004, 12(1): 37-46.

[34] Yao Y Y. Decision-Theoretic Rough Set Models[C]. *International Conference on Rough Sets and Knowledge Technology*, 2007: 1-12.

[35] 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.

[36] Qian Y H, Liang J Y, Yao Y Y, et al. MGRS: A Multi-Granulation Rough Set[J]. *Information Sciences*, 2010, 180(6): 949-970.

[37] Ziarko W. Variable Precision Rough Set Model[J]. *Journal of Computer and System Sciences*, 1993, 46(1): 39-59.

[38] Hu X H, Cercone N. Learning in Relational Databases: A Rough Set Approach[J]. *Computational Intelligence*, 1995, 11(2): 323-338.

[39] Stefanowski J, Tsoukiàs A. On the Extension of Rough Sets under Incomplete Information[C]. *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, 1999: 73-81.

[40] Yao Y Y, Zhao Y. Attribute Reduction in Decision-Theoretic Rough Set Models[J]. *Information Sciences*, 2008, 178(17): 3356-3373.

[41] Greco S, Matarazzo B, Slowinski R. Rough Sets Theory for Multicriteria Decision Analysis[J]. *European Journal of Operational Research*, 2001, 129(1): 1-47.

[42] Shu W H, Qian W B. An Incremental Approach to Attribute Reduction from Dynamic Incomplete Decision Systems in Rough Set Theory[J]. *Data & Knowledge Engineering*, 2015, 100: 116-132.

[43] Wang P, Yu Z H, Li J P, et al. Rough Set Theory and Its Application[J]. *Journal of Heze University*, 2020, 42(2): 6-10.
(王培, 于章晗, 李津平, 等. 浅谈粗糙集理论及其应用[J]. *菏泽学院学报*, 2020, 42(2): 6-10.)

[44] Zhou Z G. Research on Application of Reduction Algorithm Based on Variable Precision Rough Set in Network Educational Resource Retrieval[J]. *Information Technology and Informatization*, 2018(12): 84-85.
(周正国. 基于变精度粗糙集的约简算法在网络教育资源检索中的应用[J]. *信息技术与信息化*, 2018(12): 84-85.)

[45] Guo W. Application of Rough Set-Genetic Algorithm in Fault Diagnosis of Hydro-Generator Set[J]. *Science & Technology Vision*, 2018(35): 144-146, 154.
(郭威. 粗糙集—遗传算法在水轮发电机组故障诊断中的应用[J]. *科技视界*, 2018(35): 144-146, 154.)

[46] Li R P. *Completeness and knowledge acquisition on incomplete information system*[D]. Taiyuan: Shanxi University, 2011.
(李瑞平. 不完备信息系统的完备化及其上的知识获取[D]. 太原: 山西大学, 2011.)

[47] Slowiński R, Stefanowski J. rough-Set Reasoning about Uncertain

Data[J]. *Fundamenta Informaticae*, 1996, 27(2, 3): 229-243.

[48] Lipski W Jr. On Semantic Issues Connected with Incomplete Information Databases[J]. *ACM Transactions on Database Systems*, 1979, 4(3): 262-296.

[49] Li M, Zhang X F. Knowledge Entropy in Rough Set Theory[C]. *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, 2005: 1408-1412.

[50] Liang Jiye, Wang Baoli, Qian Yuhua, et al. A construction algorithm for maximal consistent block in an incomplete information system[C]. *CRSSC'2006,* 2006:79-82.
(梁吉业,王宝丽,钱宇华,等. 一种不完备信息系统中极大相容块的构造算法[C]. *第六届中国 Rough 集与软计算学术研讨会(CRSSC'2006)*. 2006:79-82.)

[51] Leung Y, Li D Y. Maximal Consistent Block Technique for Rule Acquisition in Incomplete Information Systems[J]. *Information Sciences*, 2003, 153: 85-106.

[52] Pang J F, Liang J Y. Granular Essence in Incomplete Decision Table[J]. *Computer Development & Applications*, 2006, 19(2): 9-11.
(庞继芳, 梁吉业. 不完备决策表中的粒度思想[J]. *电脑开发与应用*, 2006, 19(2): 9-11.)

[53] Qian Y H, Liang J Y, Wu W Z Z, et al. Information Granularity in Fuzzy Binary GrC Model[J]. *IEEE Transactions on Fuzzy Systems*, 2011, 19(2): 253-264.

[54] Wei W, Liang J Y. Information Fusion in Rough Set Theory: An Overview[J]. *Information Fusion*, 2019, 48: 107-118.

[55] Hedar A R, Wang J E, Fukushima M. Tabu Search for Attribute Reduction in Rough Set Theory[J]. *Soft Computing*, 2008, 12(9): 909-918.

[56] Miao D Q, Zhao Y, Yao Y Y, et al. Relative Reducts in Consistent and Inconsistent Decision Tables of the Pawlak Rough Set Model[J]. *Information Sciences*, 2009, 179(24): 4140-4150.

[57] Parthaláin N, Shen Q, Jensen R. A Distance Measure Approach to Exploring the Rough Set Boundary Region for Attribute Reduction[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(3): 305-317.

[58] Thangavel K, Pethalakshmi A. Dimensionality Reduction Based on Rough Set Theory: A Review[J]. *Applied Soft Computing*, 2009, 9(1): 1-12.

[59] Zhou T, Lu H L, Ren H L, et al. Survey on Attribute Reduction Algorithm of Rough Set[J]. *Acta Electronica Sinica*, 2021, 49(7): 1439-1449.
(周涛, 陆惠玲, 任海玲, 等. 基于粗糙集的属性约简算法综述[J]. *电子学报*, 2021, 49(7): 1439-1449.)

[60] Kryszkiewicz M. Rough Set Approach to Incomplete Information Systems[J]. *Information Sciences*, 1998, 112(1/2/3/4): 39-49.

[61] Kryszkiewicz M. Rules in Incomplete Information Systems[J]. *Information Sciences*, 1999, 113(3/4): 271-292.

[62] Zheng F, Wu Y Z, Hang X S. Research on Roughness of Knowledge in Rough Sets Theory[J]. *Computer Engineering and Applications*, 2002, 38(4): 98-101.
(郑芳, 吴云志, 杭小树. 粗集理论中知识的粗糙性研究[J]. *计算机工程与应用*, 2002, 38(4): 98-101.)

[63] Ma J, Gao W, Mitra P, et al. Detecting rumors from microblogs with recurrent neural networks[C]. *Proceedings of the 25th International Joint Conference on Artificial Intelligence,* 2016: 3818-3824.

[64] Ma J, Gao W, Wong K-F. Detect rumors in microblog posts using propagation structure via kernel learning[C]. *Association for Computational Linguistics,* 2017: 708-717.

王标 于 2012 年在解放军信息工程大学电子工程专业获得学士学位。现在战略支援部队信息工程大学网络空间安全专业攻读硕士学位。研究兴趣包括: 虚假舆情检测、图神经网络、粗糙集理论。Email: wangbiao9911@163.com。

卫红权 于 2014 年在解放军信息工程大学通信与信息系统专业获得博士学位。现任国家数字交换系统工程技术研究中心研究员。研究领域为融合网络安全、可重构网络理论与技术。Email: whq@ndsc.com.cn。

王凯 于 2019 年在战略支援部队信息工程大学军事信息学专业获得博士学位。现为国家数字交换系统工程技术研究中心副研究员。研究领域为网络安全治理、通信网络安全。Email: wangkai0508@126.com。

江昊聪 于 2016 年在英国布里斯托大学无线通信与信号处理专业获得硕士学位。现为国家数字交换系统工程技术研究中心助理研究员。研究领域为通信网络安全。Email: nancllia@163.com。

刘树新 于 2016 年在解放军信息工程大学获得博士学位。现为国家数字交换系统工程技术研究中心助理研究员。研究领域为复杂网络, 链路预测, 通信网络安全。Email: liushuxin11@126.com。