ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 1,Jan. 2024,pp. 181－192
DOI:10. 16337/j. 1004-9037. 2024. 01. 016

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于邻域量化容差条件熵增量式更新的网络入侵检测方法

骆公志，侯若娴

(南京邮电大学管理学院,南京 210003)

摘　要：网络入侵检测系统是网络信息安全防护的重要防御工具,而复杂的、冗长的网络入侵行为特征严重影响了网络入侵检测的效果。针对网络入侵检测中信息量增长迅速、数据不完备的现实问题,提出一种基于邻域量化容差条件熵增量式更新的特征选择算法。首先,在邻域量化容差粒计算的基础上,结合条件熵在刻画特征不确定性、对特征之间的相关或依赖程度方面的显著特性,研究了邻域量化容差条件熵的增量式更新机制;然后,基于该更新机制提出动态数据库增量式更新的特征选择算法;最后,通过数据实验分析验证了所提出的算法能有效提高不完备信息系统特征选择的计算效率。新提出的算法在网络入侵检测实例应用中体现的计算复杂度及虚警率低的优势,表明其可为网络信息安全防护提供有效可行的具体方法。

关键词：不完备信息系统;邻域粗糙集;条件熵;增量式学习;网络入侵检测

中图分类号：TP181　　**文献标志码**:A

Network Intrusion Detection Method Based on Incremental Updating of Neighborhood Valued Tolerance Condition Entropy

LUO Gongzhi, HOU Ruoxian

(School of Management, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Network intrusion detection system is an important defense tool for network information security protection, and the complicated and lengthy network intrusion behavior features seriously affect the effectiveness of network intrusion detection. In order to solve the problem of rapid information growth and incomplete data in network intrusion detection, an incremental feature selection algorithm based on neighborhood valued tolerance condition entropy is proposed. Firstly, on the basis of neighborhood valued tolerance granular computing, combined with the remarkable characteristics of conditional entropy in characterizing the uncertainty of features and the correlation or dependency between features, the incremental updating mechanism of neighborhood valued tolerance conditional entropy is studied. Then, based on the update mechanism, an incremental feature selection algorithm for dynamic database is proposed. Finally, the experimental analysis shows that the proposed algorithm can effectively improve the computational efficiency of feature selection in incomplete information systems. The new algorithm has the advantages of low computational complexity and low false alarm rate in the application of network intrusion detection examples, which shows that it can provide effective and feasible concrete methods for network

基金项目:国家自然科学基金(72171124);江苏高校哲学社会科学研究重大项目(2021SJZDA129);江苏省研究生科研创新计划项目(KYCX21_0838)。
收稿日期:2022-11-07;**修订日期**:2023-06-08

information security protection.

Key words: incomplete information system; neighborhood rough set; conditional entropy; incremental learning; network intrusion detection

引　　言

互联网的飞速发展,人类社会已经迈进了大数据时代,开放的虚拟世界伴随海量数据信息,使得网络安全问题成为社会各界持续关注的热点。世界各地频繁出现的网络攻击事件数量的急剧上升,对不同国家的众多领域的信息安全造成了不同程度的影响。如在2022年6月,中国西北工业大学的学校电子邮件系统遭受境外黑客组织网络攻击;2022年7月,挪威国家数据网络遭大规模分布式拒绝服务攻击,导致其国内公共与私营网站瘫痪数小时,在线服务停摆。网络攻击频次的大规模增长促使了网络入侵检测(Network instrusion detection, NID)的产生,NID已成为网络保护基础设施的重要组成部分。然而随着网络环境变得越来越复杂,其规模大、动态更新迅速的特征愈加明显,网络攻击模式不断演进,使得高效率、高精度的NID系统开发面临以下挑战。首先,由于数据集规模大,以及数据处理过程繁杂,常导致难以满足NID实时性的要求,而数据冗余又会带来NID误报率偏高;其次,大部分NID数据集训练是以大量的完备数据为基础来增强模型训练效果,但现实的网络数据在传输以及收集处理的多个过程中难免会出现信息丢失或数据收集不完全的情况,比如常见的出于考虑保护用户隐私安全的必要,而隐藏部分数据造成的数据不完备[1-3]。

鉴于粗糙集理论在数据处理过程中无需先验知识,以及在决策信息获取时可以降维提速的特点,众多学者尝试从粗糙集的角度研究网络入侵检测以降低其计算复杂度[4-8]。Prasad等[9]提出一种基于特征选择的贝叶斯和粗糙集入侵检测方法,对CICIDS2017数据集使用贝叶斯粗糙集提取重要特征并进行排序,删除冗余特征并有效降低了计算复杂度。Liu等[10]提出基于GA-GOGMM的模式学习和基于模糊粗糙集的属性选择的自适应入侵检测方法,基于模糊粗糙集理论,通过预先确定信息增益比,得到网络连接记录的最优属性子集,此方法有效地避免了聚类数的经验初始化和聚类中心的随机选择所造成的负面影响。但以上所提出方法的特征选择过程计算复杂度仍较高,且不能随数据库进行实时动态更新。增量式学习[11]是基于已获取的知识,针对动态增加的数据进行知识更新,其中基于粗糙集的增量式特征选择已成为从大规模动态数据集中快速提取知识的重要方法之一。在这一研究领域主要存在两种观点:(1)基于信息表的观点,由于信息表由对象、属性和数据属性值组成,因此研究主要集中在对象的变化、属性的变化以及数据属性值的变化;(2)基于拓扑学,把动态性分为两个方面,同步动态性(知识随时间演化)和历时动态性(从一个观点改变到另一个观点)[12-13]。Ciucci[14]列举了研究粗糙集动态性的4条主线,分别是上下近似集、特征选择和约简规则、精度指标和形式逻辑,这为粗糙集的增量式特征选择方法研究提供了基础框架。

在实际应用中不完备信息随处可见,且对象的动态变化显著影响知识更新[15-20],对此研究者提出了一系列有效的增量式学习算法用以不完备信息系统的知识更新并提高计算效率。Liu等[21]提出了一种基于矩阵的动态不完备信息系统增量式方法,在不完备信息系统中引入4种不同扩展关系(容差关系、相似关系、有限容差关系和特征关系)下的3个矩阵(支持矩阵、精度矩阵和覆盖矩阵)以动态更新知识,但样本间关系刻画较为宽松,容易导致误分类现象产生。Ge等[22]针对混合不完备决策系统(Hybrid incomplete decision systems, HIDS)中数据的动态变化,研究对象和属性的多层次、多维变化下概率近似的增量式更新理论和方法,提出了一种基于归一化组合关系的概率粗糙集模型,设计了基于矩阵的概率近似增量式更新算法,避免了静态算法的重复计算,但动态系统的决策值是固定不变的,这在实际动态更新系统中显然不合理。针对上述问题,在邻域量化容差条件熵的基础上,为解决不完备信息系统

中数据的动态变化,提出增量式特征选择,并通过数据实验验证了算法的有效性。

1 基本理论

设完备信息系统$IS=\{U,A\}$,其中U为对象的有限非空集,A是对象属性的有限非空集。对于$\forall a\in A$,都有映射$a:U\to V_a$,其中V_a是属性a的值集。关于属性$B\subseteq A$的子集,不可分辨关系$\mathrm{IND}(B)$可定义为:$\mathrm{IND}(B)=\{(x,y)\in U\times U:a(x)=a(y),\forall a\in B\}$。$\mathrm{IND}(B)$是不可分辨关系且$\mathrm{IND}(B)=\bigcap_{a\in B}\mathrm{IND}\{a\}$。对于任意$B\subseteq A$,不可分辨关系$\mathrm{IND}(B)$构成论域$U$的分区,由$U/\mathrm{IND}(B)$或$U/B$表示,$U/\mathrm{IND}(B)=\{[x]_B|x\in U\}$,$[x]_B$表示由$\mathrm{IND}(B)$诱导的包括$x$的等价类。对于任意子集$x\in U$,$X$的上下近似定义为:$\underline{B}X=\bigcup\{[x]_B:[x]_B\subseteq U\}$ $\overline{B}X=\bigcup\{[x]_B:[x]_B\cap X\neq\varnothing\}$。若$\overline{B}X=\overline{B}X$,$X$是可定义的,否则$X$是不可定义的。

设至少存在一个属性$a\in A$,若V_a包含缺失值,缺失值用*表示,那么$\mathrm{IIS}=\{U,A\}$被称为不完备信息系统,其中$*\in\bigcup_{a\in A}V_a$。特别地,当$A=C\cup\{D\}$时,IIS被称为不完备决策信息系统,记为$\mathrm{IDS}=\{U,A,V,F\}$,其中C为条件属性,D为决策属性,其中f为信息函数,满足映射关系$f:U\times A\to V$。

2 邻域量化容差条件熵的增量式特征选择

以下在邻域量化容差关系[23-24]的基础上,针对不完备信息系统中数据的动态变化,提出增量式特征选择。

2.1 邻域量化容差条件熵的增量式更新

定理1(划分类的增量式更新) 对于不完备决策信息系统$\mathrm{IDS}=\{U,A,V,f\}$,设$U=\{x_1,x_2,\cdots,x_n\}$,属性集$B\subseteq C$,邻域半径为δ。属性集B在信息系统中的邻域量化容差关系为NVT_B^U,得到的邻域量化容差类为

$$U/\mathrm{NVT}_B^U=\{\mathrm{NVT}_B^U(x_1),\mathrm{NVT}_B^U(x_2),\cdots,\mathrm{NVT}_B^U(x_n)\} \tag{1}$$

当新的对象集$\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+k}\}$加入该信息系统时,设新的信息系统更新为$\mathrm{IDS}'=(U',A,V,f)$,其中,$U'=U\bigcup\Delta U$。设$\mathrm{NVT}_B^{U'}$为新的信息系统确定的邻域量化容差关系,则$\forall x\in U'$的邻域量化容差关系为

$$\mathrm{NVT}_B^{U'}(x)=\mathrm{NVT}_B^U(x)\cup\left\{y\middle|x\in\mathrm{NVT}_B^{U'}(x),\forall y\in\Delta U\right\} \tag{2}$$

同时$\forall x\in U'$的新的邻域量化容差类更新为

$$U'/\mathrm{NVT}_B^{U'}=\{\mathrm{NVT}_B^{U'}(x_1),\mathrm{NVT}_B^{U'}(x_2),\cdots,\mathrm{NVT}_B^{U'}(x_n),\mathrm{NVT}_B^{U'}(x_{n+1}),\mathrm{NVT}_B^{U'}(x_{n+2}),\cdots,\mathrm{NVT}_B^{U'}(x_{n+k})\} \tag{3}$$

证明 对于任意新加入对象$y\in\Delta U$,若对象$x\in U$且$x\in\mathrm{NVT}_B^{U'}(y)$,则表明$\mathrm{NVS}_B^{\delta}(x,y)\geqslant\lambda$,即$y\in\mathrm{NVT}_B^{U'}(x)$,又$y\notin\mathrm{NVT}_B^U(x)$,所以$\mathrm{NVT}_B^{U'}(x)=\mathrm{NVT}_B^U(x)\cup\{y\}$,因此有$\mathrm{NVT}_B^{U'}(x_n)=\mathrm{NVT}_B^U(x_n)\cup\{y\big|x\in\mathrm{NVT}_B^{U'}(x),\forall y\in\Delta U\}$。

定理1给出了当不完备决策信息系统中新加入对象集$\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+k}\}$时,邻域量化容差关系和邻域量化容差类所发生的更新。以下将在此基础上给出邻域量化容差条件熵仅一个新元素加入时、以及多个新元素加入时的增量式更新。

定理2(条件熵的增量式更新) 对于不完备决策信息系统$\mathrm{IDS}=\{U,A,V,f\}$,设$U=\{x_1,x_2,\cdots,x_n\}$,属性集$B\subseteq C$,邻域半径δ。属性集B在论域U诱导出的邻域量化容差粒化为

$$U/N_B^U=\{n_B^U(x_1),n_B^U(x_2),\cdots,n_B^U(x_n)\} \tag{4}$$

当包含k个对象的新对象集$\Delta U=\{x_{n+1},x_{n+2},\cdots,x_{n+k}\}$加入信息系统后,$U'=U\cup\Delta U$,新的信息

系统 $\mathrm{IDS}'=(U',C\cup D)$,属性集 B 在论域 U' 诱导的邻域量化容差粒化为

$$U'/N_B^{U'}=\{n_B^{U'}(x_1),n_B^{U'}(x_2),\cdots,n_B^{U'}(x_{n+k})\} \tag{5}$$

决策属性 D 关于属性集 B 的邻域粒化容差条件熵为 $\mathrm{NVTE}^{U'}(D|B)$,则有

$$\mathrm{NVTE}^{U'}(D|B)=\frac{n^2}{(n+k)^2}\mathrm{NVTE}^{U}(D|B)+\frac{2}{(n+k)^2}\left(\left|\mathrm{NVT}_B^{U}(x_{n+1})-[x_{n+1}]_D^{U_1}\right|\right)+$$
$$\left|\mathrm{NVT}_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}\right|+\cdots+\left|\mathrm{NVT}_B^{U_n}(x_{n+k})-[x_{n+k}]_D^{U_n}\right| \tag{6}$$

证明 由定义5,有邻域量化容差条件熵

$$\mathrm{NVTE}^{U}(D|B)=\mathrm{NVTE}^{U}(D,B)-\mathrm{NVTE}^{U}(B)=\frac{1}{n}\left(\frac{\left|\mathrm{NVT}_B^{U}(x_i)\right|}{n}-\frac{\left|\mathrm{NVT}_B^{U}(x_i)\cap[x_i]_D^{D}\right|}{n}\right)=$$
$$\frac{1}{n^2}\sum_{i=1}^{n}\left(\left|\mathrm{NVT}_B^{U}(x_i)\right|-\left|\mathrm{NVT}_B^{U}(x_i)\cap[x_i]_D^{U}\right|\right) \tag{7}$$

当加入1个对象 x_{n+1} 时,设 $U'=U\cup\{x_{n+1}\}$,令 $\Phi=\mathrm{NVT}_B^{U'}(x_{n+1})$,$\Omega=U'-\mathrm{NVT}_B^{U'}(x_{n+1})$,根据定理1可知 Ω 表示增加对象 x_{n+1} 后论域 U 中邻域量化容差粒不发生变化的对象集,可以推出如果 $x\in\Omega$,$\mathrm{NVT}_B^{U'}(x)=\mathrm{NVT}_B^{U}(x)$,同时 $\mathrm{NVT}_B^{U'}(X)\cap[x]_D^{U'}=\mathrm{NVT}_B^{U}(X)\cap[x]_D^{U}$,则有

$$\mathrm{NVTE}^{U'}(D|B)=\frac{1}{(n+1)^2}\left(\sum_{i=1}^{n+1}\left|\mathrm{NVT}_B^{U'}(x_i)\right|-\sum_{i=1}^{n+1}\left|\mathrm{NVT}_B^{U'}(x_i)\cap[x_i]_D^{U'}\right|\right)=$$
$$\frac{1}{(n+1)^2}\left(\sum_{x\in\Omega}\left|\mathrm{NVT}_B^{U}(x_i)\right|+\sum_{x\in\Phi}\left|\mathrm{NVT}_B^{U'}(x_i)\right|-\sum_{x\in\Omega}\left|\mathrm{NVT}_B^{U}(x_i)\cap[x]_D^{U}\right|-\sum_{x\in\Phi}\left|\mathrm{NVT}_B^{U'}(x_i)\cap[x]_D^{U'}\right|\right) \tag{8}$$

若 $y\in\Phi$,则 $\mathrm{NVT}_B^{U'}(y)=\mathrm{NVT}_B^{U}(y)\cup\{x_{n+1}\}$,所以对于所有的 $y\in\Phi-\{x_{n+1}\}$,均有 $\left|\mathrm{NVT}_B^{U'}(y)\right|=\left|\mathrm{NVT}_B^{U}(y)\right|+1$,当对象 y 与新增对象 x_{n+1} 有相同决策值时,$\left|\mathrm{NVT}_B^{U'}(y)\cap[y]_B^{U'}\right|=\left|\mathrm{NVT}_B^{U}(y)\cap[y]_B^{U}\right|+1$,所以在 $\Phi-\{x_{n+1}\}$ 中有 $\left|\mathrm{NVT}_B^{U'}(x_{n+1})\cap[x_{n+1}]_B^{U'}\right|-1$ 个对象满足该条件,则有

$$\mathrm{NVTE}^{U'}(D|B)=\frac{1}{(n+1)^2}\left[\sum_{x\in\Omega}\left|\mathrm{NVT}^{U}(x)\right|+\sum_{x\in\Phi-\{x_{n+1}\}}\left(\left|\mathrm{NVT}^{U}(x)\right|+1\right)+\left|\mathrm{NVT}^{U'}(x_{n+1})\right|\right]-$$
$$\sum_{x\in\Phi-\{x_{n+1}\}}\left|\mathrm{NVT}^{U'}(x)\cap[x]_D^{U'}\right|-\left|\mathrm{NVT}^{U'}(x_{n+1})\cap[x_{n+1}]_D^{U'}\right|=\frac{1}{(n+1)^2}\left(\sum_{i=1}^{n}\left|\mathrm{NVT}^{U}(x)\right|+|\Phi|-1+\right.$$
$$\left.\left|\mathrm{NVT}^{U'}(x_{n+1})\right)\right)-\sum_{i=1}\left|\mathrm{NVT}^{U}(x_i)\cap[x_i]_D^{U}\right|-\left(\left|\mathrm{NVT}^{U'}(x_{n+1})\cap[x_{n+1}]_D^{U'}\right|-1\right)-$$
$$\left|\mathrm{NVT}^{U'}(x_{n+1})\cap[x_{n+1}]_D^{U'}\right|=\frac{1}{(n+1)^2}\left(\sum_{i=1}^{n}\left|\mathrm{NVT}^{U}(x)\right|-\sum_{i=1}\left|\mathrm{NVT}^{U}(x_i)\cap[x_i]_D^{U}\right|\right)+\left|\mathrm{NVT}^{U'}(x_{n+1})\right|+$$
$$\left|\mathrm{NVT}^{U'}(x_{n+1})\right|-\left|\mathrm{NVT}^{U'}(x_{n+1})\cap[x_{n+1}]_D^{U'}\right|-\left|\mathrm{NVT}^{U'}(x_{n+1})\cap[x_{n+1}]_D^{U'}\right|=$$
$$\frac{1}{(n+1)^2}\left(n^2\mathrm{NVTE}^{U}(D|B)+2\left|\mathrm{NVT}^{U'}(x_{n+1})\right|-2\left|\mathrm{NVT}^{U'}(x_{n+1})\cap[x_{n+1}]_D^{U'}\right|\right) \tag{9}$$

根据集合的运算关系:$\left|\mathrm{NVT}^{U'}(x_{n+1})\right|-\left|\mathrm{NVT}^{U'}(x_{n+1})\cap[x_{n+1}]_D^{U'}\right|=\left|\mathrm{NVT}^{U'}(x_{n+1})\right|-\left|[x_{n+1}]_D^{U'}\right|$,

所以$\mathrm{NVTE}^{U'}(D|B)=\frac{1}{(n+1)^2}\left(n^2\mathrm{NVTE}^{U}(D|B)+2\left|\mathrm{NVT}^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}\right|\right)$。

当加入多个对象$x_{n+1},x_{n+2},\cdots,x_{n+k}$时,设$U'=U\cup U_k$,$U_k=\{x_{n+1},x_{n+2},\cdots,x_{n+k}\}$,$\mathrm{NVT}_B^{U'}(x_{n+i})$为对象$x_{n+i}$在论域$U_i$上的邻域量化容差粒;$[x_{n+i}]_D^{U_i}$为对象$x_{n+i}$在论域$U_i$上的决策类。

$$\mathrm{NVT}_B^{U_3}(D|B)=\frac{(n+2)^2}{(n+3)^2}\mathrm{NVT}_B^{U_2}(D|B)+\frac{2\left|\mathrm{NVT}_B^{U_3}(x_{n+3})-[x_{n+3}]_D^{U_3}\right|}{(n+3)^2}=$$

$$\frac{(n+2)^2}{(n+3)^2}\left[\frac{n^2}{(n+2)^2}\mathrm{NVT}_B^{U}(D|B)+\frac{2\left|\mathrm{NVT}_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}\right|}{(n+2)^2}+\frac{2\left|\mathrm{NVT}_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}\right|}{(n+2)^2}\right]+$$

$$\frac{2\left|\mathrm{NVT}_B^{U_3}(x_{n+3})-[x_{n+3}]_D^{U_2}\right|}{(n+3)^2}=\frac{n^2}{(n+3)^2}\mathrm{NVT}_B^{U}(D|B)+\frac{2\left|\mathrm{NVT}_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}\right|}{(n+3)^2}+$$

$$\frac{2\left|\mathrm{NVT}_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}\right|}{(n+3)^2}+\frac{2\left|\mathrm{NVT}_B^{U_3}(x_{n+3})-[x_{n+3}]_D^{U_3}\right|}{(n+3)^2}\tag{10}$$

依次计算有

$$\mathrm{NVT}_B^{U_i}(D|B)=\frac{n^2}{(n+i)^2}\mathrm{NVT}_B^{U}(D|B)+\frac{2}{(n+i)^2}\left(\mathrm{NVT}_B^{U_1}(x_{n+1})+\mathrm{NVT}_B^{U_1}(x_{n+2})\right)-$$
$$\left([x_{n+2}]_D^{U_2}+\cdots+\mathrm{NVT}_B^{U_i}(x_{n+i})-[x_{n+i}]_D^{U_i}\right)\tag{11}$$

所以,$\mathrm{NVT}_B^{U_k}(D|B)=\mathrm{NVT}_B^{U'}(D|B)=\frac{n^2}{(n+k)^2}\mathrm{NVT}_B^{U}(D|B)+\frac{2}{(n+k)^2}(\mathrm{NVT}_B^{U_1}(x_{n+1})-[x_{n+1}]_D^{U_1}+\mathrm{NVT}_B^{U_2}(x_{n+2})-[x_{n+2}]_D^{U_2}+\cdots+\mathrm{NVT}_B^{U_k}(x_{n+k})-[x_{n+k}]_D^{U_k})$。

定理2表明,当不完备信息系统加入新对象时,只需要依次计算出每个新加入对象的邻域量化容差粒和决策类。对象x_{n+1}在$U\cup\{x_{n+1}\}$上进行计算,对象x_{n+2}在$U\cup\{x_{n+1},x_{n+2}\}$上进行计算,x_{n+i}在$U\cup\{x_{n+1},x_{n+2},\cdots,x_{n+i}\}$上进行计算,在计算过程中,可以逐步计算邻域量化容差粒,多个对象依次加入信息系统,加入一个对象时便立即在当时的信息系统内计算邻域量化容差粒和决策类,当所有的对象加入完毕,便可计算整个邻域量化容差条件熵。

2.2 邻域量化容差条件熵增量式特征选择算法

由定理2可知:$\mathrm{NVTE}^{U'}(D|B)=\frac{n^2}{(n+1)^2}\mathrm{NVTE}^{U}(D|B)+\frac{2\left|\mathrm{NVT}_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}\right|}{(n+1)^2}$。

对两个邻域量化容差条件熵相减,即

$$\mathrm{NVTE}^{U'}(D|B)-\mathrm{NVTE}^{U}(D|B)=\frac{n^2}{(n+1)^2}\mathrm{NVTE}^{U}(D|B)+\frac{2\left|\mathrm{NVT}_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}\right|}{(n+1)^2}-$$
$$\mathrm{NVTE}^{U}(D|B)=\frac{2\left|\mathrm{NVT}_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}\right|}{(n+1)^2}-\frac{2n+1}{(n+1)^2}\mathrm{NVTE}^{U}(D|B)\tag{12}$$

由定义5,有$0\leqslant\mathrm{NVTE}^{U}(D|B)\leqslant1-1/n$,令$q=\left|\mathrm{NVT}_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}\right|$,经计算得:$2(q-n)+\frac{1}{n}+1\leqslant\mathrm{NVTE}^{U'}(D|B)-\mathrm{NVTE}^{U}(D|B)\leqslant2q$。由于$\mathrm{NVT}_B^{U'}(x_{n+1})-[x_{n+1}]_D^{U'}\subseteq U$,所以$0\leqslant$

$q \leqslant n$,有 $-2n+1/n+1 \leqslant 2(q-n)+1/n+1 \leqslant 1/n+1$,由此可推,$2(k-n)+1/n+1$ 的值可能为正、负或0,$\mathrm{NVTE}^{U'}(D|B)-\mathrm{NVTE}^{U}(D|B)$ 的值也是如此。所以当信息系统增加1个对象后,其邻域量化容差粒的变化是不确定的。

初始信息系统的特征选择采用算法(Feature selection based on neighborhood valued tolerance conditional entropy, FSNVTCE)[23],当不完备信息系统有新对象加入时,传统的非增量式特征选择算法要直接对新的信息系统进行处理,随着数据集的规模不断扩大,特征选择的计算复杂度会越来越大,而提出的增量式特征选择算法采用增量式的学习方法,在原信息系统特征选择的基础上,增量式地计算信息系统的特征选择,大幅度提高计算效率,具体的增量式学习算法步骤如下。

算法 基于邻域量化容差条件熵的增量式特征选择算法

输入:不完备决策信息系统 $\mathrm{IDS}=\{U, C\cup D\}$,$|U|=n$,邻域半径 δ,邻域相似度阈值 λ,终止阈值 $\varepsilon=0.000\,1$;IDS 中条件属性集合 C 的约简集 RED;邻域量化容差条件熵 $\mathrm{NVTE}^{U}(D|C)$,$\mathrm{NVTE}^{U}(D|\mathrm{RED})$,新增对象 $x_{n+1}, x_{n+2}, \cdots, x_{n+k}$。

输出:IDS′中的特征选择结果 RED′。

(1)初始化 $\mathrm{NVTE}^{U'}(D|C)=1$,$\mathrm{NVTE}^{U'}(D|\mathrm{RED})=1$。

(2)根据 $\mathrm{NVTE}^{U}(D|C)=1$ 和 $\mathrm{NVTE}^{U}(D|\mathrm{RED})=1$,按定理2增量式计算 $\mathrm{NVTE}^{U'}(D|C)$ 和 $\mathrm{NVTE}^{U'}(D|\mathrm{RED})$。

(3)计算 $\mathrm{sig}(a,B,D)$,其中 $\mathrm{sig}(a,B,D)=\mathrm{NVTE}(D|B)-(D|B\cup\{a\})$。若 $\mathrm{sig}(a,B,D)>\varepsilon$,则跳转至步骤(4);若 $\mathrm{sig}(a,B,D)\leqslant\varepsilon$,则转至步骤(7)。

(4)对于 $\forall a\in C-\mathrm{RED}$,计算每个条件属性对应的邻域量化容差条件熵 $\mathrm{NVTE}^{U'}(D|\mathrm{RED}\cup\{a\})$。

(5)选择满足下列条件的属性,记为 a^*:$\max\limits_{\forall a\in C-\mathrm{RED}}[\mathrm{NVTE}^{U'}(D|\mathrm{RED})-\mathrm{NVTE}^{U'}(D|\mathrm{RED}\cup\{a\})]$

(6)若 $\mathrm{sig}(a^*,B,D)>\varepsilon$,则 $\mathrm{RED}=\mathrm{RED}\cup\{a^*\}$,并转至步骤(4),否则转至步骤(7)。

(7)$\mathrm{RED}'=\mathrm{RED}$,返回约简集合 RED′。

在IFSNVTCE的步骤(2)中,初始 $\mathrm{NVTE}^{U}(D|C)$ 与 $\mathrm{NVTE}^{U}(D|\mathrm{RED})$ 的值均为1,按照定理2增量式更新 $\mathrm{NVTE}^{U'}(D|C)$ 与 $\mathrm{NVTE}^{U'}(D|\mathrm{RED})$。根据信息系统论域增大,邻域量化容差条件熵的变化必定满足 $\mathrm{NVTE}^{U'}(D|C)\leqslant\mathrm{NVTE}^{U'}(D|\mathrm{RED})$,因此在步骤(3)需要判断 $\mathrm{sig}(a,B,D)$ 的大小,若 $\mathrm{sig}(a,B,D)\leqslant\varepsilon$,则算法终止,返回约简集RED;若 $\mathrm{sig}(a,B,D)>\varepsilon$,则需要在剩余的条件属性中继续搜索。

邻域量化容差条件熵增量式特征选择算法(Incremental of feature selection based on neighborhood valued tolerance conditional entropy, IFSNVTCE)是在FSNVTCE基础上提出的,增量式特征选择算法每次运算只需要计算新增的数据量,而静态特征选择算法需要对新增数据量后的整个数据集进行计算。对算法进行计算复杂度分析可知,IFSNVTCE为贪婪算法,在算法的执行过程中每次选取剩余属性中属性重要度最高的属性,再将其加入约简集RED中。设最终约简集RED包含 r 个属性。设 $|C\cup D|=q$,$|\mathrm{RED}|=r$,IFSNVTCE的时间复杂度为 $O(rm\log_2 n)$,当 $r=q$ 时为耗时最多的情形,此时的时间复杂度为 $O(qn\log_2 n)$,计算复杂度明显低于非增量式算法的复杂度。

3 数据实验

3.1 数据标准化处理

在进行数据处理之前,要先对数据集进行数据标准化,数据标准化是进行数据挖掘与知识发现的重要步骤,数据集中的每个值都要缩放到合适的范围内,这个过程有利于消除特征偏差,数据标准化公

式表示为

$$x_{ij}=[x_{ij}-\min(x_j)]/[\max(x_j)-\min(x_j)] \tag{13}$$

式中:x_{ij}为样本值的标准化值,其取值范围为0～1;$\min(x_j)$为第j个属性的最小值;$\max(x_j)$为第j个属性的最大值。

为了对IFSNVTCE算法性能进行测试,从UCI数据库中获取6个标准数据集进行实验,数据集描述如表1所示。在每个数据集中构造属性值缺失比例不同的数据集,以此来评估算法在不完备信息系统中的特征选择性能。

表1 实验数据集

Table 1 Experimental dataset

序号	数据集	对象数	特征数	类别数
1	Iris	150	8	3
2	Car	270	13	2
3	Raisin	900	7	2
4	Abalone	4 177	8	3
5	Wdbc	208	60	2
6	Move	360	90	15

3.2 IFSNVTCE与FSNVTCE时间消耗比较

IFSNVTCE算法的性能分析在Windows 11操作系统,i5-11400h,CPU 2.7 GHz, RAM 16 GB硬件平台,采用Matlab 2020a编程环境完成。

设定邻域半径δ从0.03～0.15以步长0.03递增取值,相似度阈值λ从0.1～0.6以步长0.1取值,每组取值都会得到对应的特征选择结果,将特征选择后的数据集送入SVM、C4.5分类器并采用十折交叉法计算分类精度,最终选取分类精度较高时所对应的约简结果,各数据集的特征选择结果如表2所示。

表2 各数据集特征选择结果

Table 2 Feature selection results of each dataset

数据集	缺失比例/%	(λ,δ)	特征选择个数	特征选择结果	SVM/%	C4.5/%
Iris	10	(0.1,0.03)	4	5,8,2,1	91.3	92.2
	20	(0.1,0.09)	7	5,8,2,1,4,3,7	89.6	90.5
Car	10	(0.1,0.03)	6	1,2,6,4,5,3	92.1	91.7
	20	(0.2,0.06)	8	1,2,6,4,8,5,7,3	90.0	89.5
Raisin	10	(0.3,0.03)	3	3,2,5	82.3	83.4
	20	(0.3,0.06)	5	3,1,2,5,7	79.7	78.2
Abalone	10	(0.2,0.06)	4	4,8,6,1	93.2	92.7
	20	(0.1,0.09)	6	1,4,8,2,6,3	91.3	92.5
Wdbc	10	(0.3,0.06)	7	28,22,27,13,5,23	92.4	90.6
	20	(0.1,0.06)	10	28,22,27,13,5,23,3,9,16,19	88.1	91.4
Move	10	(0.1,0.12)	14	12,45,33,26,18,34,11,7,5,14,32,46, 70,19	78.3	81.5
	20	(0.1,0.15)	22	12,45,4,33,29,21,35,26,17,43,18 34,37,11, 7,39,5,14,32,46,70,19	77.9	82.3

两种算法的特征选择结果是相同的,通过将这两种算法对表2中数据集的特征选择消耗时间进行比较,用时较少的即为更优的算法。图1、2为在数据集缺失比例分别为10%和20%的情况下两种算法在各个数据集特征选择的时间消耗比较,为了构造数据集对象的动态增加,将整个数据集大致分为10等份,随机选取1份作为初始数据集,然后从其他每份中随机选取1份对象集加入数据集中,模拟出数据集9次动态增加的情形。

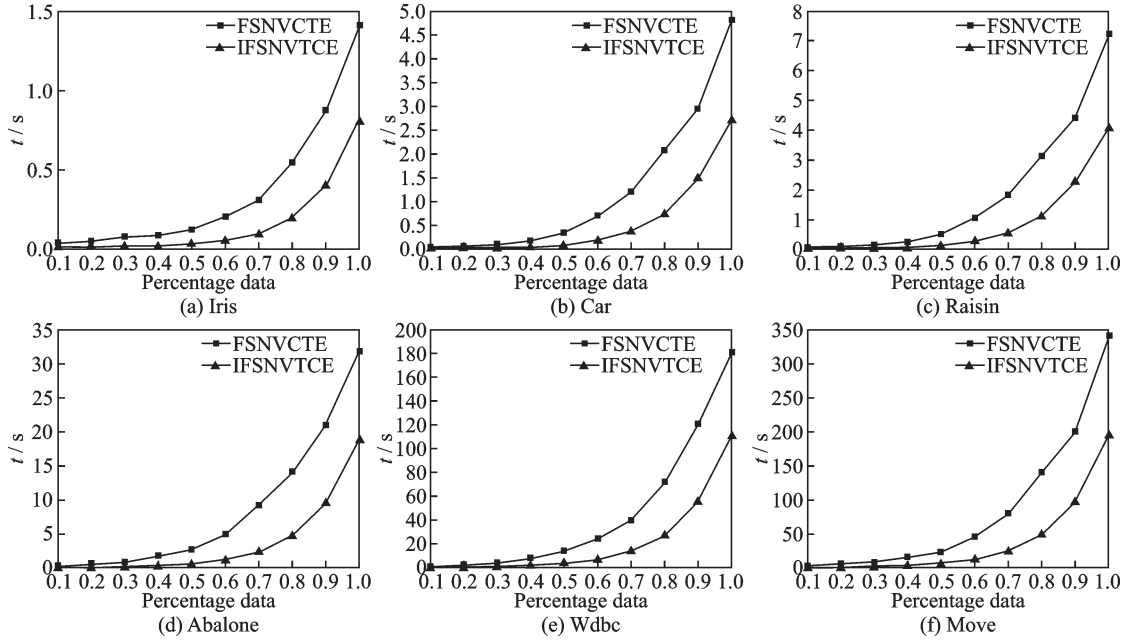

图1 数据集缺失比例10%时IFSNVTCE和FSNVTCE特征选择的时间消耗对比

Fig.1 Comparison of time consumption for feature selection between IFSNVTCE and FSNVTCE with the datasets 10% missing

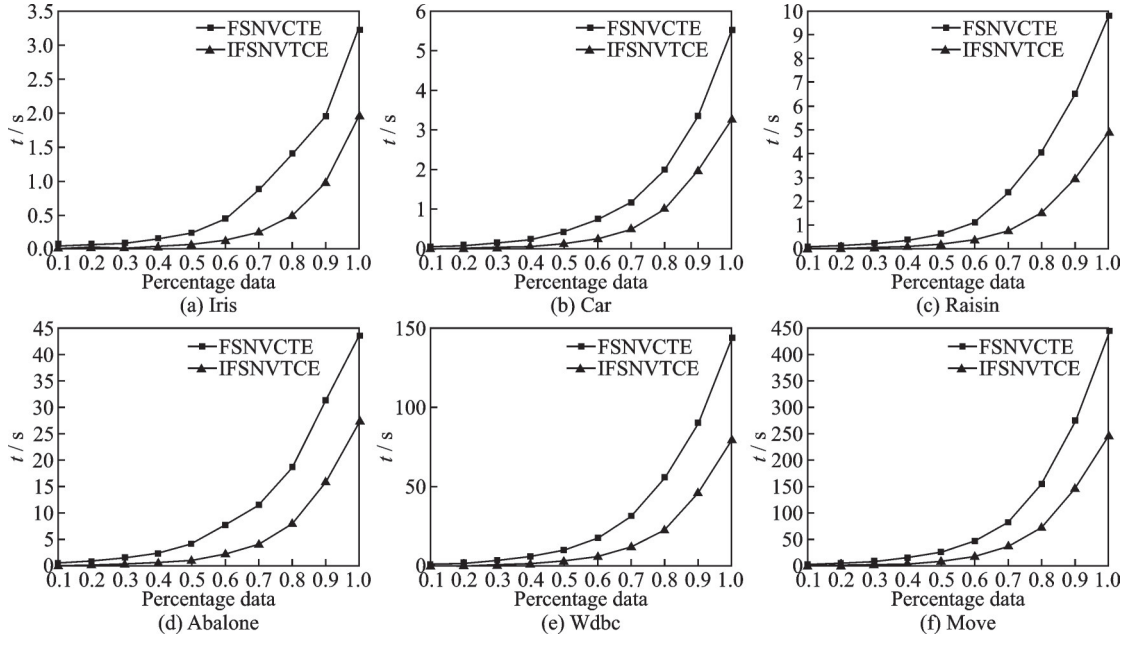

图2 数据集缺失比例20%时IFSNVTCE和FSNVTCE特征选择的时间消耗对比

Fig.2 Comparison of time consumption for feature selection between IFSNVTCE and FSNVTCE with the datasets 20% missing

由图1、2可知，当数据集开始按比例动态增加后，两种算法特征选择用时发生明显变化，IFSNVTCE有效降低了算法的计算复杂度。

4 实例应用及结果分析

以下选取标准NSL-KDD数据集来验证IFSNVTCE算法在实际NID中的应用性能。

4.1 NSL-KDD数据集

NSL-KDD数据集解决了经典网络入侵数据集KDD99的一些固有问题,如KDD99数据集中最重要的缺陷是训练集和测试集中存在大量冗余记录(在训练集和测试集中分别各有约为78%和75%的记录重复),这会导致分类器更偏向于频繁出现的样本记录,从而妨碍分类器学习稀疏攻击类型的样本记录,但U2R和R2L等不频繁记录对网络安全危害更大。相比之下,NSL-KDD数据集中的训练集和测试集有合理数量的记录且没有冗余数据,其中包括训练集125 973条记录和测试集22 544条记录,因此分类器不会偏向于更频繁的记录,可用于NID系统实验[25-26]。NSL-KDD数据集包含41个条件属性,可分为3类:基本连接属性、网络流量特征和网络内容特征,决策属性是样本的攻击类型,分为4类:DoS(Denial of service attack)、Probe、R2L(Remote-to-login attack)和U2R(User-to-root attack)。DoS攻击是拒绝服务攻击,目的是让目标网络无法提供正常的服务,使目标系统停止响应甚至崩溃,最常见的DoS攻击有计算机网络带宽攻击和连通性攻击;Probe攻击是一种网络攻击,指监视或其他探测,常见于端口扫描;R2L攻击是来自远程机器的未授权访问,多用于密码猜测;U2R攻击是指未经授权的用户绕过一些系统或者网站的漏洞直接获取最高权限,然后登陆进行非法操作,如各种缓冲区溢出攻击。攻击类型具体如表3所示。

表3 NSL-KDD数据集中的网络攻击类型

Table 3 Network attack types in NSL-KDD dataset

类别	描述	攻击类型
Benign	正常网络流量	normal
DoS	拒绝服务攻击	apache2, back, mailbomb, processtable, snmpgetattack, teardrop, smurf, land, Neptune, pod, udpstorm
Probe	远程用户攻击	ftp_write, guess_passwd, sumpguess, imap, spy, warezclient, warezmaster, multihop, phf, named, sendmail, xlock, xsnoop
R2L	收集网络信息	nmap, ipsweep, portsweep, satan, mscan, saint, worm
U2R	通过非法手段获得最高权限	Ps, buffer_overflow, perl, rootkit, loadmodule, xterm, sqlattack, httptunnel

4.2 属性选择效果

通过使用NSL-KDD数据集中提供的两个不同训练集来验证特征选择方法的结果,即完整的训练集NSL-KDD Train和它的子集NSL-KDDTrain_20,以及两个不同的测试集,完整的测试集NSL-KDD Test和它的子集KDDTest-21,训练集和测试集的组成如表4所示。

表4 NSL-KDD数据集结构

Table 4 Structure of NSL-KDD dataset

数据集	样本总数	正常样本数量	攻击样本数量
NSL-KDDTrain_20	25 192	13 448	11 743
NSL-KDD Train	125 973	67 342	58 630
KDDTest-21	11 850	2 152	9 697
NSL-KDD Test	22 544	9 711	12 833

特征选择有利于降低后续网络异常检测的时间损耗。网络入侵数据集的数值缺失比例可能会因数据集而异,但一般情况下,缺失比例会在5%到20%之间[27]。将NSL-KDD数据集分别随机构造10%、20%的缺失比例,图3、4分别为在NSL-KDD数据集两种缺失比例状态下FSNVT-

CE算法与IFSNVTCE算法特征选择的时间消耗比较。同样,为了构造数据集对象的动态增加,不妨将整个数据集大致分为10等份,随机选取一份作为初始数据集,然后从其他每份中随机选取一份对象集加入数据集中,模拟出数据集9次动态增加的情形,数据集特征选择结果如表5所示。

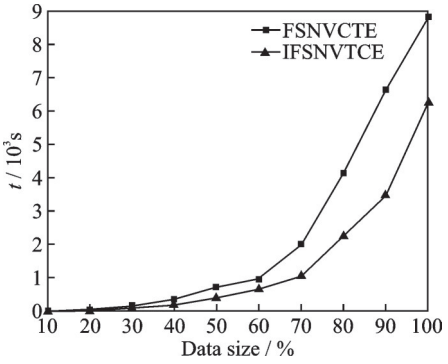

图3 NSL-KDD数据集缺失比例10%时IFSNVTCE和FSNVTCE特征选择的时间消耗对比

Fig.3 Comparison of time consumption for feature selection between IFSNVTCE and FSNVTCE with NSL-KDD dataset 10% missing

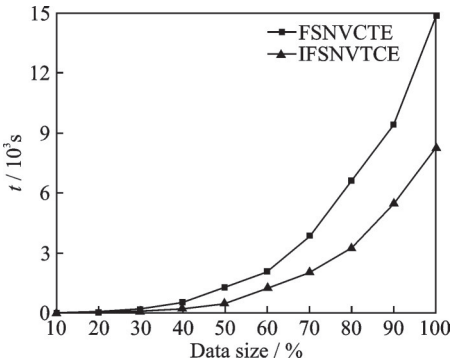

图4 NSL-KDD数据集缺失比例20%时IFSNVTCE和FSNVTCE特征选择的时间消耗对比

Fig.4 Comparison of time consumption for feature selection between IFSNVTCE and FSNVTCE with NSL-KDD dataset 20% missing

表5 NSL-KDD数据集特征选择结果

Table 5 NSL-KDD dataset feature selection results

类别	缺失比例/%	(λ,δ)	特征选择个数	特征选择结果	SVM/%	C4.5/%
NSL-KDD Train_20	10	(0.1, 0.06)	20	1,2,3,5,6,11,15,17,18,20,21,22,23,25,29,31,33,34,35,39	93.3	92.7
	20	(0.2, 0.09)	29	1,3,4,5,6,9,11,12,13,15,17,18,19,20,21,22,23,25,26,28,29,31,33,34,35,37,38,39,41	91.5	92.0
KD-DTest-21	10	(0.1, 0.03)	23	2,5,6,9,11,12,13,15,17,18,19,20,21,25,26,27,28,33,34,35,37,38,39	93.6	89.5
	20	(0.2, 0.06)	26	3,4,5,6,9,10,11,12,13,15,16,17,18,19,20,21,23,25,26,28,33,34,35,37,38,39	90.7	92.3

综上所述,基于IFSNVTCE的入侵检测方法能在数据不完备的情况下保持较高的检测精度。该方法的优点主要来自基于IFSNVTCE的最优属性选择以及正常和入侵网络连接记录模式学习的在线更新。IFSNVTCE可以去除原始实例中的冗余、不确定信息。此外,所引入的增量式更新策略可以保证网络入侵检测能够适应网络环境的改进和变化。因此,对于已知攻击和不可预见攻击,提出的网络入侵检测方法可以在时间消耗较低的情况下实现较高的检测精度。

5 结束语

特征选择是粗糙集理论研究的核心问题,其计算复杂度是决定粗糙集模型运行效率的主要因素。通过提出一种基于不完备信息系统增量式特征选择的网络入侵检测方法,对网络入侵数据集进行特征选择,选择最优属性子集。先对于现实中样本信息不完备、信息缺失的情况提出了相应解决方法;然后

对于真实网络环境下网络样本增长快、特征选择过程耗时长的问题,提出增量式特征选择算法。数据实验分析表明,相对于传统的特征选择方法,本文提出的算法对能够有效降低计算复杂度、在提取高维数据集的特征数据时可以显著缩短运行时间。网络入侵检测数据分析的应用实例,验证了本文算法在处理网络入侵数据时的有效性和实用性,可广泛用于实际的网络入侵检测中,为网络安全防护提供科学的理论依据与具体的实践路径。

参考文献:

[1] LIN S W, YING K C, LEE C Y, et al. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection[J]. Applied Soft Computing, 2012, 12(10): 3285-3290.

[2] KHRAISAT A, GONDAL I, VAMPLEW P, et al. Survey of intrusion detection systems: Techniques, datasets and challenges[J]. Cybersecurity, 2019, 2(1): 1-22.

[3] OTHMAN S M, BA-ALWI F M, ALSOHYBE N T, et al. Intrusion detection model using machine learning algorithm on big data environment[J]. Journal of Big Data, 2018, 5(1): 1-12.

[4] 刘文博,梁盛楠,董小刚.基于t类加权核函数的主成分分析维度约简算法[J].统计与决策,2022,38(9): 52-56.
LIU Wenbo, LIANG Shengnan, DONG Xiaogang. PCA dimension reduction algorithm based on t class weighted kernel function[J]. Statistics & Decision, 2022, 38(9): 52-56.

[5] 李金海,王飞,吴伟志,等.基于粒计算的多粒度数据分析方法综述[J].数据采集与处理,2021,36(3): 418-435.
LI Jinhai, WANG Fei, WU Weizhi, et al. Review of multi-granularity data analysis methods based on granular computing[J]. Journal of Data Acquisition and Processing, 2021, 36(3): 418-435.

[6] MOHEIMANI A, SHEIKH R, HOSSEINI S M H, et al. Assessing the preparedness of hospitals facing disasters using the rough set theory: Guidelines for more preparedness to cope with the COVID-19[J]. International Journal of Systems Science: Operations & Logistics, 2022, 9(3): 339-354.

[7] LUO J, QIN K, ZHANG Y, et al. Incrementally updating approximations based on the graded tolerance relation in incomplete information tables[J]. Soft Computing, 2020, 24(12): 8655-8671.

[8] MANDAL P, RANADIVE A S. Multi-granulation interval-valued fuzzy probabilistic rough sets and their corresponding three-way decisions based on interval-valued fuzzy preference relations[J]. Granular Computing, 2019, 4(1): 89-108.

[9] PRASAD M, TRIPATHI S, DAHAL K. An efficient feature selection based Bayesian and rough set approach for intrusion detection[J]. Applied Soft Computing, 2020, 87: 105980.

[10] LIU J, ZHANG W, TANG Z, et al. Adaptive intrusion detection via GA-GOGMM-based pattern learning with fuzzy rough set-based attribute selection[J]. Expert Systems with Applications, 2020, 139: 112845.

[11] SU H, QI W, HU Y, et al. An incremental learning framework for human-like redundancy optimization of anthropomorphic manipulators[J]. IEEE Transactions on Industrial Informatics, 2020, 18(3): 1864-1872.

[12] 陈海燕,刘晨晖,孙博.时间序列数据挖掘的相似性度量综述[J].控制与决策,2017,32(1): 1-11.
CHEN Haiyan, LIU Chenhui, SUN Bo. Survey on similarity measurement of time series data mining[J]. Control and Decision, 2017, 32(1): 1-11.

[13] 张文冬,亓慧,刘克宇,等.基于粗糙集特征选择的过拟合现象及应对策略[J].南京航空航天大学学报,2019,51(5): 687-692.
ZHANG Wendong, QI Hui, LIU Keyu, et al. Over-fitting and its countermeasure in feature selection based on rough set[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2019, 51(5): 687-692.

[14] CIUCCI D. Classification of dynamics in rough sets[C]//Proceedings of International Conference on Rough Sets and Current Trends in Computing. Berlin, Heidelberg: Springer, 2010: 257-266.

[15] SHEN H, DAI M, LUO Y, et al. Fault-tolerant fuzzy control for semi-Markov jump nonlinear systems subject to incomplete SMK and actuator failures[J]. IEEE Transactions on Fuzzy Systems, 2020, 29(10): 3043-3053.

[16] WANG G, GUAN L, WU W, et al. Data-driven valued tolerance relation based on the extended rough set[J]. Fundamenta Informaticae, 2014, 132(3): 349-363.

[17] WAN R, MIAO D, PEDRYCZ W. Constrained tolerance rough set in incomplete information systems[J]. CAAI Transactions

on Intelligence Technology, 2021, 6(4): 440-449.

[18] WANG X Y, ZHU T, SHEN Y X. Research on the application of limited tolerance relation in multi granularity rough set[C]// Proceedings of 2021 4th International Conference on Algorithms, Computing and Artificial Intelligence. New York, United States: Association for Computing Machinery, 2021: 1-5.

[19] GUAN L. Data-driven valued dominance relation in incomplete ordered decision system[J]. Knowledge and Information Systems, 2021, 63(11): 2901-2917.

[20] ARUNKUMAR C, RAMAKRISHNAN S. Prediction of cancer using customised fuzzy rough machine learning approaches[J]. Healthcare Technology Letters, 2019, 6(1): 13-18.

[21] LIU D, LI T, ZHANG J. A rough set-based incremental approach for learning knowledge in dynamic incomplete information systems[J]. International Journal of Approximate Reasoning, 2014, 55(8): 1764-1786.

[22] GE H, YANG C. Incremental updating probabilistic approximations under multi-level and multi-dimensional variations in hybrid incomplete decision systems[J]. International Journal of Approximate Reasoning, 2022, 142: 206-230.

[23] 姚晟,徐风,赵鹏,等.基于邻域量化容差关系粗糙集模型的特征选择算法[J].模式识别与人工智能, 2017, 30(5): 416-428.
YAO Sheng, XU Feng, ZHAO Peng, et al. Feature selection algorithm based on neighborhood valued tolerance relation rough set model[J]. Pattern Recognition and Artificial Intelligence, 2017, 30(5): 416-428.

[24] ZHAO H. Intrusion detection ensemble algorithm based on bagging and neighborhood rough set[J]. International Journal of Security and Its Applications, 2013, 7(5): 193-204.

[25] 张晓琴,汪云飞,胡春强.基于改进极限学习机的数据采集与监控系统攻击检测模型[J].南京航空航天大学学报, 2021, 53(5): 708-717.
ZHANG Xiaoqin, WANG Yunfei, HU Chunqiang. Attack detection model of SCADA system based on data preprocessing and improved ELM[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(5): 708-717.

[26] DEVI R R, ABUALKIBASH M. Intrusion detection system classification using different machine learning algorithms on KDD-99 and NSL-KDD datasets—A review paper[J]. International Journal of Computer Science and Information Technology, 2019, 11(3): 65-80.

[27] RING M, WUNDERLICH S, SCHEURING D, et al. A survey of network-based intrusion detection data sets[J]. Computers & Security, 2019, 86: 147-167.

作者简介:

骆公志(1972-),男,博士,教授,研究方向:粗糙集理论及应用,E-mail: lgzlyg@163.com。

侯若娴(1997-),**通信作者**,女,硕士研究生,研究方向:粗糙集理论及应用,E-mail: houruoxiansd@163.com。

(编辑:刘彦东)