网络传输态势感知的研究与实现

卓莹，龚春叶，龚正虎
（国防科学技术大学 计算机学院，湖南 长沙 410073）

摘 要：将态势感知的先进思想引入网络传输领域，以空间流量聚类为基本思想，建立网络传输态势感知（NTSA）模型；围绕模型关键技术，依据信息增益和互信息的等价性执行态势因子选择，提出了一种面向传输模式划分的高维数据流聚类算法，并且基于图论进行拓扑重要性分析；设计并且实现了 NTSA 原型系统。基于真实数据集的实验验证了系统的时效性、准确性以及可扩展性。

关键词：计算机体系结构；NTSA；模型；空间流量分析；聚类；特征选择；图论
中图分类号：TP393 **文献标识码**：A **文章编号**：1000-436X(2010)09-0054-10

Research and implementation of network transmission situation awareness

ZHUO Ying, GONG Chun-ye, GONG Zheng-hu
(School of Computer Science, National University of Defense Technology, Changsha 410073, China)

Abstract: The advanced ideas of situation awareness were introduced to network transmission and NTSA (network transmission situation awareness) model was established based on spatial traffic clustering. Around the key technologies of the model, situation factors were selected according to information gain and mutual information; a high-dimensional data stream clustering algorithm for transmission pattern partition as well as a topology importance analysis method of network element based on graph theory were proposed; furthermore, a NTSA prototype system was designed and implemented. The experiment results on real datasets demonstrate the efficiency, effectiveness and scalability.

Key words: computer architecture; NTSA; model; spatial traffic analysis; clustering; feature selection; graph theory

1 引言

随着信息网络规模的迅速扩大，系统复杂性也随之增加。传统的网络管理中各功能单元处于独立的工作状态，缺少有效的信息提取和信息融合机制，无法建立网络资源之间的联系，全局信息表现能力差。海量的网管信息非但不能加强管理，反而增加了网络管理员的负担。现代网络管理必须能够提供多样化、个性化的管理行为，提供被管对象的详细信息，了解整个网络的运行状况，并且能够根据指挥人员的需求提供服务。因此，基于融合的网络态势感知必将成为网络管理的发展方向[1,2]。

传输是网络最基本的功能。网络传输态势感知（NTSA）是指在大规模信息网络中，从传统的单元网管获取链路流量信息、资源运行状况、网络拓扑结构等能够对传输产生影响的各种测量指标（称

收稿日期：2009-01-15；**修回日期**：2010-03-10
基金项目：国家重点基础研究发展计划（“973”计划）基金资助项目（2009CB320503）；国家高技术研究发展计划（“863”计划）基金资助项目（2008AA01A325）
Foundation Items: The National Basic Research and Development Program of China (973 Program) (2009CB320503); The National High Technology Research and Development Program of China (863 Program) (2008AA01A325)

为态势因子），评估当前状态、预测未来发展趋势，并以可视化的方式进行展示。NTSA 包括获取、评估、预测以及可视化 4 个环节，强调网络元素之间的关系，决定了态势因子的汇聚方式，给出一种有利或不利的判断性结果。简而言之就是态势因子集合 R 与态势空间框架 θ 的映射关系 $f:R\to\theta$。

NTSA 的目标是将态势感知技术应用于网络管理，在急剧动态变化的复杂环境中，高效组织各种信息，将已有的表示网络局部特征的指标综合化，使其能够表示网络传输的宏观整体状态，从而加强管理员对网络的理解能力，为高层指挥人员提供决策支持。

态势感知的研究主要包括 3 个方面：模型、知识表示和评估方法。其中模型研究比较完善，趋于统一；评估方法作为态势感知的重点，理论研究相对成熟[3]；有关知识表示的研究相对较少[4]。然而现有的研究面向通用问题，没有很好地将态势感知技术运用到网络传输领域，缺少针对网络传输具体特点设计的模型和评估方法；另一方面，有关网络传输的研究还没有上升到态势的高度，仍旧停留在数据层面上，采用的指标体系比较简单，没有综合分析各种态势因子，而层次结构、加权函数为主流的评估方法缺少理论支持，无法展现网络传输系统元素间错综复杂的关系。

本文充分利用态势感知研究成果，结合网络传输具体问题展开 NTSA 研究。首先结合空间流量分析和数据挖掘技术，提出了基于空间流量聚类的网络传输态势感知模型（STC 模型）。接下来分别讨论了特征选择、聚类算法、拓扑分析等关键技术：证明了互信息和信息增益的等价性，并以此为基础进行态势因子选择，降低态势空间规模；在分析网络数据特点的基础上，提出了一种面向传输模式划分的高维数据流聚类算法（SACluster 算法），并且结合聚类和粗集技术设计了网元传输态势评估方法；以图论为基础，进行拓扑重要性分析，确定网元对传输态势的影响。最后设计了 NTSA 系统逻辑结构，实验和试运行论证了系统的可行性和有效性，并且简要给出相关工作以及比较结果。

2　基于空间流量聚类的 NTSA 模型

针对现有研究存在的问题，本文结合空间流量分析和数据挖掘技术，提出了基于空间流量聚类的网络传输态势感知模型——STC 模型。首先介绍 STC 模型的基本思想。

所谓空间流量分析[5]，相对时间流量分析而言，不仅仅分析单条链路流量的时间行为，而是综合考虑链路流量特征和互连方式，分析跨越多条链路或全网链路的流量模式，支持全网视图。空间流量分析利用“所有网络事件都会在流量上有所反映”这一基本假设[6]，实现了网络拓扑和流量特征的交互，能够建立高度概括的完整的网络态势视图。

空间流量分析最大的挑战在于缺少先验知识。传输态势涉及的测量指标众多，单一指标对传输产生的影响不同，指标之间也存在复杂的关系相互影响，加之相关研究很少，没有对传输模式进行划分的成果可以借鉴。如果通过领域专家对传输模式进行划分，不可避免带有主观色彩，而且很难公平准确地判断每一个指标对态势做出的贡献。聚类作为数据挖掘的主要方法，属于无监督机器学习方法，具备获取知识、揭示规律的能力。聚类分析能够根据海量网络数据自动完成对传输态势的模式划分，提取典型态势特征，不需要任何先验知识，科学客观。

STC 模型弥补了简单指标、单条链路的不足，既能够融合多种影响传输态势的因素，揭示各种已知和未知的异常事件，避免了采用层次结构和权重分析方法的缺陷；又能够综合考虑拓扑结构的影响，揭示网络元素间错综复杂的关系，实现全网传输态势感知，体现了态势整体性和宏观性的特点；而且模式聚类扩展灵活、适用性强，不局限于流量特征。

结合空间流量数据的特点，对网络传输态势感知进行建模：*STC*(*Topo*, *Traffic*)，其中 *Topo* 代表拓扑信息，*Traffic* 代表流量信息。

拓扑信息包括网络元素及其之间的连接关系，表示为 *Topo*(*ID*, *Time*, *Node*, *Link*, ψ)。其中，*ID* 代表某一次拓扑发现的标识，是唯一的。*Time* 代表此次拓扑发现的时间。*Node* 代表节点集合，使用四元组(ID_N, *C*, *W*, *Dsc*)来表示；其中，ID_N 是节点的唯一标识；*C* 代表节点处理能力；*W* 代表节点的重要性权值，由网络拓扑结构以及节点处理能力决定；*Dsc* 用于描述节点相关信息，是一个可扩展项。*Link* 代表链路集合，与节点集合类似，统一使用四元组(ID_L,*C*,*W*,*Dsc*)来表示，不同之处在于 *C* 代表链路容量。ψ 代表连接关系，$\psi\subseteq Node\times Node$。

流量信息包括通过流量分析挖掘所获取的有关态势感知的各类信息和知识，表示为 *Traffic*(*ID*,

Time, *Trace*, *IS*, *SF*, *SP*, *AR*)。其中：*ID* 为某个网元（节点或者链路统称为网络元素，简称网元）的标识。*Time* 代表流量发生时间。*Trace* 代表流量采集获得的原始报文信息。*IS* 代表指标体系（index system），通过对 *Trace* 进行流量分析获得反映网络传输态势的特征的集合。$IS=\{a_1,a_2,\cdots,a_n\}$，其中，$a_i$ 为流量特征。*SF* 代表态势因子（situation factor），经过特征选择获得能够引起网络态势发生变化的重要因素的集合；$SF=\{f_1,f_2,\cdots,f_d\}$，其中，$f_k$ 为态势因子，$f_k \in IS$。*SF* 是指标体系的子集，$SF \subset IS$。*SP* 代表态势模式（situation pattern），依据态势因子的取值，通过聚类对传输态势模式进行划分；每种态势模式形如：

$$sp=<(f_1,v_1),(f_2,v_2),\cdots,(f_d,v_d)>$$

其中，v_k 为态势因子 f_k 的取值。*AR* 代表评估规则（assessment rule），在态势模式划分的基础上，确定每种模式的态势取值，生成态势评估规则。

基于 STC 模型进行传输态势感知建模的流程如下：

{ 对拓扑数据进行建模 }

step1 拓扑发现 TopoDiscovery，获取拓扑信息 *Topo*(*ID*, *Time*, *Node*, *Link*, ψ)；

step2 拓扑推理 TopoReference(*Node*, *Link*, ψ)，修正拓扑信息；

step3 拓扑分析 TopoAnalysis(ψ, *C*)，确定链路的重要性 W；

{ 对流量数据进行建模 }

step4 流量采集 TrafficCollection，获取原始流量数据 Trace；

step5 流量分析 TrafficAnalysis(*Trace*)，建立指标体系 IS；

step6 对每一个指标 $ai \in IS$，进行离散化 Discretization(*ai*)和标准化 Standardization(*ai*)；

step7 流量特征选择 FeatureSelection(*IS*)，建立态势因子集合 *SF*；

step8 在态势因子集合上进行模式划分 PatternMining(*Trace*,*SF*)，建立态势模式 *SP*；

step9 依据态势模式划分结果进行规则设计 RuleDesign(*SP*)，建立态势评估规则 *AR*；

{ 人机交互，专家分析和确认 }

step10 模式分析 PatternAnalysis(*SP*)；

step11 模型调整 ModelAdjustment(*W*，*AR*)；

{迭代}

step12 判断模型是否有效，如果不再适用，重复以上步骤进行建模。

其中，拓扑分析（step 3）执行拓扑重要性分析，确定网元对传输态势的影响；特征选择（step 7）执行态势因子选择，缩小指标体系的规模，建立精简的态势空间；模式挖掘（step 8）执行传输态势模式划分，通过聚类分析提取网元传输模式。下面将对上述 STC 模型的 3 个关键技术，逐一进行深入讨论。规则设计（step 9）通过引入粗集分析，自动生成态势评估规则，相关内容已经另文叙述[7]。其他技术或者来自传统的网络管理，例如拓扑发现和推理（step1 和 step2）、流量采集和分析（step4 和 step5）；或者可以借鉴成熟的研究成果，例如数据标准化和离散化（step 6）；或者需要领域专家的参与，例如模式分析和模型调整（step10 和 step11），本文不作介绍。

3 关键技术

3.1 基于 IG/MI 的态势因子选择

降低数据空间的维度主要有 2 种方法。①特征提取方法，或者称为维度约简，通过线性或者非线性组合原始维度产生新的维度，把原始数据空间映射到使用新维建立的低维空间，最具代表性的是主成分分析 PCA。②特征选择方法，即从维度空间中选择一个维度子集，往往需要用户指定。由于流量指标都有明确的意义，揭示了流量的某种特征；但是新维的含义则很难给出解释[8]。考虑到评估结果易于理解，本文选择后者。

常用的特征选择方法包括文档频率(DF, document frequency)、信息增益(IG, information gain)和互信息(MI, mutual information)等，在文本分类中得到深入研究并广泛应用[9]。由于态势评估尽可能采用便于监测的指标，又经过 level 1 融合和 level 2 数据标准化，因此数据比较完整，使得 DF 失去作用，本文只讨论 IG 和 MI。

不同于文本分类，一个特征只区分出现与否 2 种情况；高维数据的每一维可以有多个不同的取值，需要综合所有取值对分类的作用。为此，对 IG 和 MI 的定义进行如下修改。

定义 1 设态势模式 *S* 共有 *n* 种，记作 $S_k(k=1, 2, \cdots, n)$。令 $S=\sum_{k=1}^{n} S_k$ 表示样本总数，则模式 S_k 的概率用 $p(S_k)=S_k/S$ 估计，简记为 p_k。为简单起见，没有区分态势模式的名称和数量$|S_k|$，统一记作 S_k。

态势因子 F 共有 d 个，对于某一个因子 $F_i(i=1, 2, \cdots, d)$ 可能有 m 种取值，记作 $v_j(j=1, 2, \cdots, m)$。则互信息定义为：

$$MI(S,F)=\sum_{s,f}p(s\wedge f)\log\frac{p(s\wedge f)}{p(s)p(f)}$$

信息增益定义为：$IG(F)=I(S)-E(F)$。

尽管已有的实验结果表明：IG 是最有效的特征选择方法之一；DF 的效果稍差，但和 IG 基本相似；而 MI 相对较差。然而笔者发现，考虑特征取值的影响对原有定义进行修改之后，IG 和 MI 2 种方法对于特征的评价是一致的。下面给出 IG 和 MI 等价性的证明。

证明

$$\begin{aligned}&IG(F)=I(S)-E(F)\\&=-\sum_{k=1}^{n}p_k\mathrm{lb}p_k+\sum_{j=1}^{m}p_j\sum_{k=1}^{n}p_{k|j}\mathrm{lb}p_{k|j}\\&=-\sum_{k=1}^{n}\sum_{j=1}^{m}p_{kj}\mathrm{lb}p_k+\sum_{j=1}^{m}\sum_{k=1}^{n}p_jp_{k|j}\mathrm{lb}p_{k|j}\\&=-\sum_{k=1}^{n}\sum_{j=1}^{m}p_{kj}\mathrm{lb}p_k+\sum_{j=1}^{m}\sum_{k=1}^{n}p_j\frac{p_{kj}}{p_j}\mathrm{lb}\frac{p_{kj}}{p_j}\\&=\sum_{k=1}^{n}\sum_{j=1}^{m}p_{kj}\mathrm{lb}\frac{p_{kj}}{p_kp_j}\\&=\sum_{s,f}p(s\wedge f)\mathrm{lb}\frac{p(s\wedge f)}{p(s)p(f)}=MI(S,F)\end{aligned}$$

故 IG 和 MI 是等价的。其中，$p_j=p(v_j)=v_j/S$，$p_{kj}=p(S_k\wedge v_j)$。

无论是信息增益还是互信息，都是衡量 2 个随机变量相互之间独立程度的测度，反映了态势因子 f 导致态势模式 S 不确定度的缩减量。由定义可知，*IG*/*MI* 的取值越大，f 和 S 之间的相关性越强，f 越重要。由于缺少态势模式的先验知识，态势因子选择只能在聚类的基础上，计算 F 和 S 之间的 *IG*/*MI*。

3.2　面向传输模式划分的聚类算法

考虑到网络传输的特点，对聚类算法提出了更高的要求，具体分析如下。① 网络数据是一种典型的数据流，数据量大，潜在无限，到达速率不确定，只能按到达顺序访问且仅能被扫描一次或有限几次，因此要求聚类算法满足一遍扫描，有限内存，输入顺序不敏感等原则。② 网络测量数据具有连续性、动态变化，要求聚类算法既能够跟上流的速度，又能够反映流的演化情况。③ 指标体系庞大、维数众多，要求聚类算法可以解决高维数据聚类问题，并且具有良好的可扩展性。④ 高维数据的稀疏性，导致数据在低维子空间形成聚类，而在高维空间没有聚类特征。⑤ 网络测量数据是结构化数据，测量指标既有连续型又有离散型，要求聚类算法可以处理混合属性数据。

根据以上分析，提出了一种面向传输模式划分的聚类(SACluster)算法。SACluster 算法首先对数据空间进行“网格”划分，在此基础上，进行全空间聚类，通过合并相连密集网格形成簇。紧接着进行子空间聚类，对不满足密度阈值的簇采用自顶向下的策略、兼顾密度与维度双重标准进行 2 次聚类：通过降低聚类空间的维度，使不满足密度阈值的簇实现相连，从而建立所有可能包含投影簇的子空间；然后进行迭代，在候选子空间中逐一搜索最优投影簇，直到发现所有满足密度阈值的投影簇，并且采用滑动窗口技术实现增量更新，动态维护聚类结果。所谓最优投影簇，是指簇中的点尽可能多，簇的维度尽可能高，即投影簇质量函数取值最大。

SACluster 算法描述如下：

step1　聚类空间网格划分；

step2　一遍扫描数据流，统计网格密度信息；

step3　合并相连密集网格（密度大于 τ_g），在全空间建立簇；

step4　输出满足密度阈值 τ_c 的簇；

step5　合并任意 2 个密度满足[τ_s,τ_c)的簇，建立最高维度候选投影簇；

step6　搜索不满足密度阈值 τ_c 的簇，统计候选投影簇的密度信息；

step7　从候选投影簇中选出一个最优投影簇；

step8　如果最优投影簇满足密度阈值 τ_c，输出该投影簇；

step9　重复 step 6~step 8，直到没有满足密度阈值 τ_c 的投影簇或者达到终止条件；

step10　如果到达更新间隔，增量更新聚类结果。

SACluster 算法结合密度和网格 2 种方法，有效降低大规模数据流聚类空间的规模，扩展灵活，能够处理混合属性数据，产生任意形状的簇，且对噪声和输入顺序不敏感。子空间方法使用原始维而非新维建立子空间，简单、易理解，同时有效解决了高维数据稀疏性问题。自顶向下的搜索策略充分利用网络数据的分布特征，满足数据流一遍扫描的需求，而且实现了在不同维度的不同子空间搜索投影簇。增量更新既能够反映数据流的演化过程；又以

较短的时间间隔更新结果，满足在线聚类对响应时间的要求。借助滑动窗口技术，有效降低算法复杂性的同时，保证参与聚类的样本量，维护模式划分稳定性。增量更新算法已经另文叙述[10]。

SACluster 是一个高效的高维数据流子空间聚类算法。假设有 n 个数据点，分布在 g 个网格中，则全空间聚类的时间复杂度为 $O(n+g^2)$；簇个数 c，不满足密度阈值 τ_c 的簇个数近似取 c（略小于 c），候选子空间的个数 s，投影簇的个数 p，则子空间聚类的时间复杂度为 $O(c^2+pcs)$。可以通过限制参与聚类的密集网格的密度 τ_g 以及建立子空间的候选簇的密度 τ_s 来降低算法复杂度。如果态势因子 d 个，增量更新维护 w 个窗口，则算法的空间复杂度为 $O(wdg)$。

3.3 基于图论的拓扑重要性分析

评估整个网络的传输态势，还需要了解每个网元对网络传输态势的影响，即链路/节点的拓扑重要性，主要包括网元对网络拓扑的贡献以及网元自身的传输能力 2 个方面。本文运用图论中有关“容量网络”的理论[11]，进行网元的拓扑重要性分析。

在图论中，网络是指具有 2 个特定顶点：发点（source）x 和收点（sink）y 的加权连通图，记作 $N=(D_{xy}, w)$。若 N 为非负的容量函数 c，则称网络 $N=(D_{xy}, c)$为容量网络（capacity network）。对应到传输网络，图论中的“容量网络”是指网络中 2 个节点：源（source）和目的（destination）之间所有路径（path）组成的包含传输能力信息的连接图。为了避免混淆，下文中全部采用传输网络术语。假设传输网络 $N(node, link, \psi, c)$，由 v 个节点（node）、ε 条链路（link）组成，ψ 表示节点到链路的关联函数（incident function），c 表示网元传输能力，即容量函数；则网络中包含 r 种连接，$r=v(v-1)/2$，即任意 2 个节点之间的连接，记作 R_{xy}。

首先讨论链路的重要性。对于源 s 到目的 d 的连接 R_{sd}，当连接中的所有链路正常运行时，该连接可以达到最大传输容量 $C_N(R_{sd})$；若某一链路 l 失效，最大传输容量将受到影响而减小，记作 $C_{N-\{l\}}(R_{sd})$。$C_N(R_{sd})$和 $C_{N-\{l\}}(R_{sd})$的差别反映了链路 l 对连接 R_{sd} 的影响。由此定义链路 l 对连接 R_{sd} 的重要性 $LI_{l,R_{sd}}$：

$$LI_{l,R_{sd}}=1-\frac{c_{N-\{l\}}(R_{sd})}{c_N(R_{sd})}=\frac{c_N(R_{sd})-c_{N-\{l\}}(R_{sd})}{c_N(R_{sd})} \quad (1)$$

进而定义链路 l 对整个传输网络的重要性 LI_l：

$$LI_l=\sum_{i=1}^{r}LI_{l,R_i}\Big/r=\sum_{i=1}^{r}\frac{c_N(R_i)-c_{N-\{l\}}(R_i)}{c_N(R_i)}\Big/r \quad (2)$$

其中 $l=1, 2, \cdots, \varepsilon$。式（2）$C_N(R_i)$的计算建立在网络拓扑结构的基础上，充分考虑了链路 l 的拓扑重要性。分母 $C_N(R_i)-C_{N-\{l\}}(R_i)$表示链路 l 为连接 R_i 提供的传输能力，以链路容量为依据，体现了链路对传输的贡献。

接下来讨论节点的重要性。节点重要性同样考虑拓扑重要性和处理能力 2 个方面。若节点失效，则以该节点为顶点的链路也会失效，可见节点的重要性受到与之关联的所有链路重要性的影响，前者应该是后者重要性之和。同时考虑到节点处理能力 t 和关联链路传输容量存在差别，若前者小于后则，则该节点将会成为瓶颈限制网络传输能力。据此分析定义容量因子 F_n：

$$F_n=\begin{cases}1, & t_n\geqslant\sum_{l=1}^{d}c_l\\ t_n\Big/\sum_{l=1}^{d}c_l, & t_n<\sum_{l=1}^{d}c_l\end{cases} \quad (3)$$

其中，d 为节点 n 的度，即与节点关联的链路的数目；$\sum_{l=1}^{d}c_l$ 表示关联链路传输容量。

则节点重要性 NI_n 定义如下：

$$NI_n=F_n\sum_{l=1}^{d}LI_l,\ n=1, 2, \cdots, v \quad (4)$$

式（4）中 F_n 反映节点处理能力对传输的影响，而链路拓扑重要性以及节点的度隐含地体现了节点的拓扑重要性。

分析了链路/节点重要性之后，只需对重要性进行简单的归一化处理（令 $\sum_l w_l+\sum_n w_n=1$），即可获得网元重要性权值。

$$\begin{cases}w_l=\dfrac{LI_l}{\sum_{i=1}^{v}NI_i+\sum_{j=1}^{\varepsilon}LI_j}\\ w_n=\dfrac{NI_n}{\sum_{i=1}^{v}NI_i+\sum_{j=1}^{\varepsilon}LI_j}\end{cases} \quad (5)$$

图论中计算最大传输容量的方法比较成熟，例如标号法[12,13]。对网元传输能力 c 为非负整数的整容量网络，标号法是有效算法，复杂度为 $O(v\varepsilon^2)$。

4　系统设计与实验分析

本文在 STC 模型的基础上，结合网络传输这一具体应用，提出了网络传输态势感知系统结构。如图 1 所示，NTSA 体系结构坚持 closing-the-loop 的理念，始终将人作为 NTSA 中的一个重要环节，突出动态循环的本质，强调反馈的重要作用，体现了 DFIG 模型的 6 层结构以及 Endsley 模型对态势感知的细化。该体系结构包括通信模块、知识发现模块、评估模块、预测模块、可视化模块、人机交互模块、自主管理模块以及模式表。

不同于 DFIG 模型的信息总线结构，本文采用基于 Web 服务的信息交换机制构建通信平台，作为各级融合进行信息交换和互操作的方式，实现了 NTSA 原型系统。

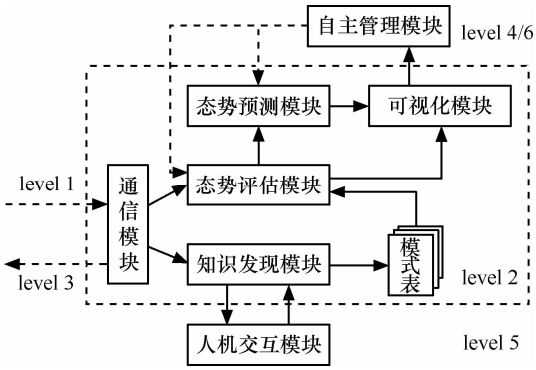

图 1　网络传输态势感知系统结构

为了验证 STC 模型，本节对所提出的算法以及原型系统进行测试。实验平台配置如下：AMD Athlon Dual Core 4200+GHz/2GB，Windows XP，所用代码均用 ActivePerl（5.8.8）实现。实验所使用的数据有 2 种，第 1 种是来源于 MIT Lincoln 实验室的 KDDCup'99 入侵检测数据集[14]，第 2 种数据来源于美国应用网络研究国家实验室(NLANR)被动测量和分析工作组(PMA)在 HPC 网络中设置多个测量点被动测量 Internet 数据[15]。

4.1　实验 1：SACluster 算法和态势因子选择

由于缺少有关网络传输态势的研究作为比较，本文采用通用的测试数据集 KDDCup'99 验证聚类 SACluster 算法。KDDCup'99 记录了 4 898 431 条流（flow）记录，分为正常模式和 22 种入侵模式，每条记录点包括 1 维连接模式（正常或者入侵）和 41 维属性，其中连续型属性 33 维，离散型属性 8 维[注 1]。KDDCup'99 数据集来自真实的网络，和本文研究的网络传输数据具有相似性，故可以用来做测试数据。

聚类准确性：本文使用聚类纯度[16]和分类正确率 2 个指标衡量聚类质量。

定义 2　聚类纯度。聚类纯度定义为簇/投影簇中主体类别 i 的数据点在簇/投影簇中所占的比例，形式化描述 $purity=\sum_{j=1}^{c}\max_{i}\left\{C_j^i\right\}\Big/\sum_{j=1}^{c}\left|C_j\right|$，$c$ 为簇的个数。

定义 3　分类正确率。分类正确率定义为在参与聚类的数据点中最终被正确划分到真实分类的比例，形式化描述如下，n 为数据点总数。

$$preciseness=\frac{1}{n}\sum_{j=1}^{c}\underset{p}{preciseness}\left\{C_j^p\right\}$$

聚类纯度由簇纯度和投影簇纯度共同决定，如图 2 所示，簇纯度全部达到 100%，投影簇纯度随数据点个数的增加逐渐降低，但由于投影簇覆盖的数据点个数较少，对纯度的影响也较小，故纯度始终保持较高水平，均在 98%以上。

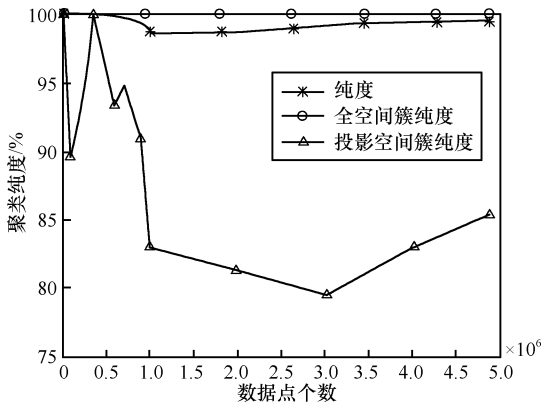

图 2　聚类纯度

由于非密集网格不参与聚类，使得非密集网格覆盖的数据点没有被划分到簇/投影簇中，因此分类正确率受到网格密度阈值的限制，不可能超过密集网格覆盖数据点的比例，如图 3 所示，τ_g=0.01%。

注 1　在 KDDCup'99 描述文档中，特征 su_attempted 的类型被标记为离散型，实际在数据集中特征 su_attempted 记录了尝试“su root”命令的次数，取值 0，1，2，应该属于连续型。

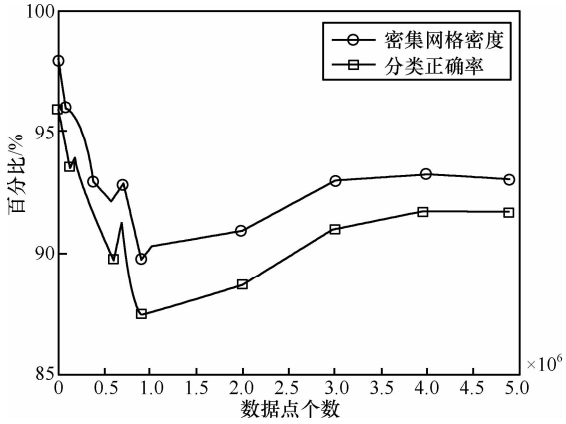

图 3 分类正确率

时间复杂性：SACluster 的初始化聚类包括统计网格（grid）分布以及合并相连网格形成簇（cluster）2 个阶段，图 4 显示了指数坐标下算法运行时间，τ_c=0.01, τ_g=0.01τ_c, τ_s=0.10τ_c。结果表明，随着数据点个数的增加，grid 时间线性增加，而 cluster 时间增长较慢。说明算法具有良好的规模可扩展性。

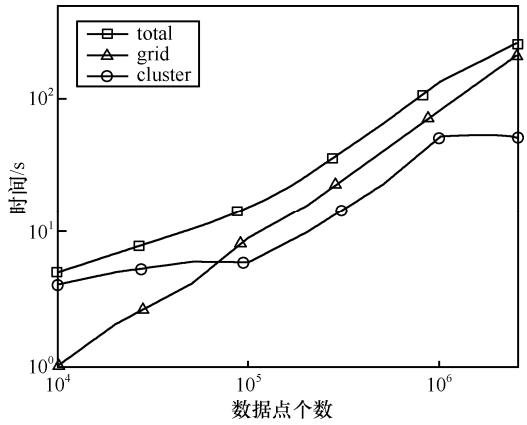

图 4 执行效率

态势因子选择的效果：从 KDDCup’99 选择 10 组数据，每组 100 000 个数据点，分别根据聚类结果以及连接模式信息计算每一维特征与记录类型之间的互信息，并对各组数据取平均值。如图 5 所示，2 条曲线极其吻合，说明根据聚类结果计算的近似互信息与已知分类情况获得的互信息一致，如实反映了特征与模式之间的相关性，显示了特征对聚类的重要程度，因此根据聚类结果获得的互信息能够作为态势因子选择的依据。同时还注意到，根据聚类结果计算的互信息偏高，这是因为在基于网格的聚类结果中去除了噪声以及低密度网格的影响，故每一维特征包含的有关模式划分的信息量增大。

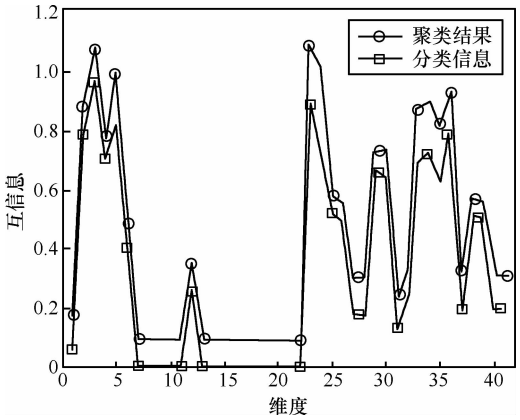

图 5 特征与类型之间的互信息

维度可扩展性：紧承上一实验，验证 SACluster 算法的维度可扩展性，实验数据保持不变。根据态势因子选择的结果，选取不同数量的特征，分别取 2、7、17、19、22、26、41 维特征。聚类运行时间如图 6 所示，仍旧分成 grid 和 cluster 2 个阶段。可以看到，随着数据点个数的增加，grid 时间线性增加，而 cluster 时间趋于平缓。说明算法具有良好的维度可扩展性。

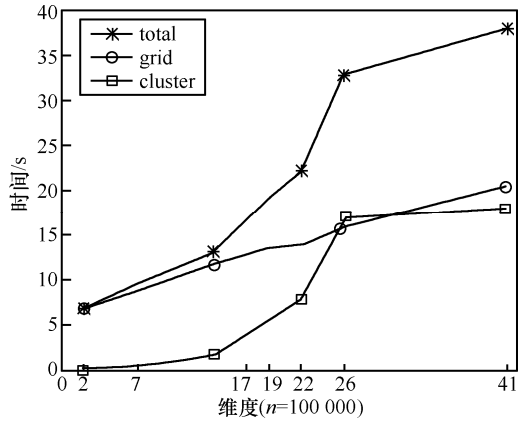

图 6 维度可扩展性

态势因子选择（维度）对聚类准确性的影响：如图 7 所示，随着聚类空间维度的增加，聚类纯度略有增加，但由于纯度始终保持较高水平，故增加不明显；分类正确率首先明显提升，继而缓慢下降，这是由于维度的增加，使得密集网格的个数有所减少，密集网格覆盖记录点的比例也随之减少，分类正确率因为受到网格密度阈值的限制有所降低。

4.2 实验 2：NTSA 原型系统

为了检验 NTSA 原型系统的效果，本文采用 NLANR 提供的 Abilene 网络流量数据。Abilene[17] 网络是美国教育科研网，如图 8 所示，其核心网络拓扑包括 11 个节点和 14 条双向链路。NLANR 数据

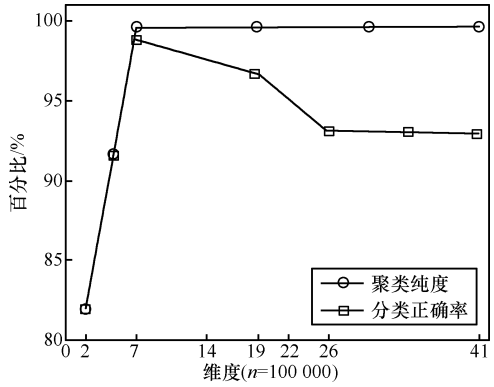

图 7　特征选择对精度的影响

图 8　Abilene 骨干网

采集 Abilene 网络的报文信息，每天采样 8 次，每次 90s。并且于 2001 年 7 月~9 月、2003 年 1 月、2004 年 1 月记录了 Code Red Worm、Slammer worm、W32 Mydoom 的爆发。

NTSA 原型系统选择了报文个数、带宽、估计延迟、报文长度分布、报文协议分布、带宽协议分布、新流个数、活动流个数、流平均时长、流平均报文数、流平均字节数、流协议分布、TCP 标志位分布等多个指标评估传输态势，图 9 显示了某链路 2 天的评估结果，时间间隔选择 1min。从图中可以看到评估结果既反映了传输状态以 24h 为单位的周期性变化，也体现出流量的异常变化。

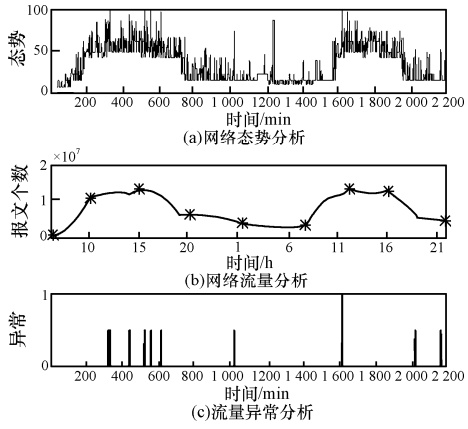

图 9　2 天态势评估

NTSA 原型系统在 Code Red Worm、Slammer worm 和 W32 Mydoom 爆发的 3 个时段进行传输态势评估。如图 10 所示，当网络异常发生时，态势曲线发生明显变化，取值相对较大。

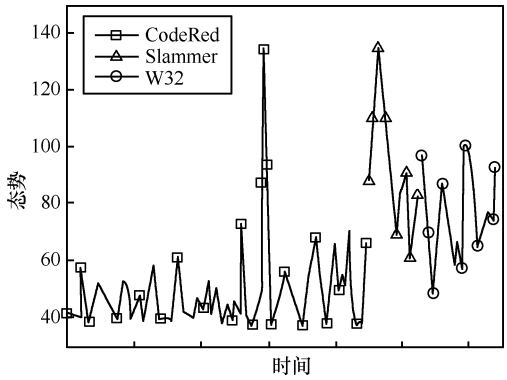

图 10　注入异常

NTSA 原型系统不仅能够很好地反映网络异常，而且能够检测未知异常并及时提取异常特征。为了全面细节地揭示异常的特征，展现异常发生时各种态势因子的表现，采用雷达图绘制态势因子图谱，如图 11 所示。在雷达图上，各种异常的特征一目了然，不仅能够同时表现多种态势因子，而且便于不同异常以及正常状态之间的比较。

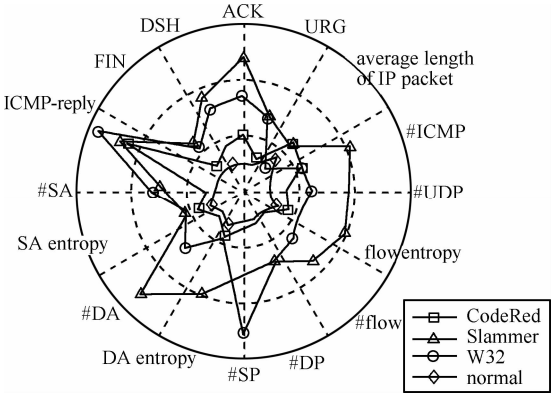

图 11　态势因子图谱

4.3　相关工作和比较

自 1999 年 Tim Bass 提出网络态势感知的概念以来，网络态势感知已经成为网络管理和网络安全领域的热点，绝大多数研究围绕安全态势展开[2,18~20]，也有少量涉及传输态势、信息优势、生存性等。有关传输态势感知的研究，集中在态势可视化阶段。Lau 等提出的 Internet 级网络流量可视化工具，在三维空间（例如选择 IP 地址和端口号建立 IP-IP-port 逻辑空间）用点来表示网络流量信息，提

表 1 相关工作比较

代表性工作		数据来源	评估方法	态势因子	结果形式
网络态势	Bass	IDS/sniffer	定性	少，固定	状态转换
	Tadda	IDS/日志/流量	公式法	少，固定	打分
安全态势	Chen	IDS/网络性能指标	层次加权，主观	少，固定	态势曲线
	Wei	各种漏洞/服务日志	证据理论，客观低效	较多，可扩展	态势曲线
传输性能	Lin	性能指标	公式法	少，固定	多目标值
	Zhang	性能/拓扑	权重分析	少，固定	目标值
流量可视化	Lau	流量	定性	少，固定	三维视图
传输态势	本文	流量/拓扑	模式匹配，客观高效	多可扩展	态势曲线，特征图谱

供整个网络的态势感知，易于发现网络攻击行为，提取攻击行为特征[21~24]。国内一些关于网络传输性能评价的研究和网络传输态势感知关系比较紧密，其中杨雅辉建立了网络性能指标体系，并给出形式化描述框架[25]，林闯、江勇等以评价函数作为标准展开网络传输控制的性能评价[26,27]，张冬艳等以权重分析为基础评价网络性能[28,29]。

本文是对传输态势感知的第一次尝试，侧重于揭示网络运行状况，结合流量和拓扑信息创新性的提出了 STM 模型。由于和现有研究差异较大，因此从数据来源、态势因子、评估方法、结果形式等几方面进行简单比较，结果见表 1。

从结果可以看出，本文方法扩展灵活、科学客观、运行效率高。安全态势感知和本文思路比较接近，不同之处在于侧重点不同，因此选择的数据源也不同。而流量可视化和性能评价等方法，或者停留在数据层面上，或者采用公式法分析少数指标，都没有上升到态势的高度。

5 结束语

本文面向网络管理需求，结合态势感知的先进技术，建立了基于空间流量聚类的网络传输态势感知模型，证明了信息增益和互信息用于态势因子选择的等价性，提出了一种面向传输模式划分的高维数据流聚类算法，设计了基于图论的拓扑重要性分析方法。实验分析证明了算法的准确性、时空可行性以及可扩展性，原型系统能够综合各种单元网管信息，从宏观上把握网络态势，体现了系统的应用价值和现实意义。本文是对网络传输态势感知的一次尝试，相关研究还有巨大的发展空间。在接下来的工作中，将关注态势感知的理论方法，深化关键技术研究，同时利用 CPU 双核的特点，进一步提高模式聚类的精度和效率。

参考文献：

[1] BASS T. Multisensor data fusion for next generation distributed intrusion detection systems[A]. 1999 IRIS National Symposium on Sensor and Data Fusion[C]. Laurel, 1999. 24-27.

[2] BASS T. Intrusion detection systems and multisensor data fusion[J]. Communications of the ACM, 2000, 43(4): 99-105.

[3] HINMAN M. Some computational approaches for situation assessment and impact assessment[A]. ISIF[C]. New York, USA, 2002. 687-693.

[4] ZHUO Y, ZHANG Q, GONG Z H. Cyberspace situation representation based on niche theory[A]. ICIA[C]. Zhangjiajie, China, 2008. 1400-1405.

[5] CROVELLA M, KOLACZYK E. Graph wavelets for spatial traffic analysis[A]. Infocom[C]. 2003. 1848-1857.

[6] LAKKARAJU K. NVisionIP: netflow visualizations of system state for security situational awareness[A]. ACM Workshop Visualization and Data Mining for Computer Security[C]. New York, USA, 2004. 65-72.

[7] ZHUO Y, ZHANG Q, GONG Z H. Network situation assessment based on RST[A]. PACIIC[C]. Wuhan, China, 2008. 502-506.

[8] AGRAWAL R, GEHRKE J, GUNOPULOS D. Automatic subspace clustering of high dimensional data for data mining applications[A]. SIGMOD[C]. 1998. 94-105.

[9] 徐燕,李锦涛,王斌.基于区分类别能力的高性能特征选择方法[J].软件学报,2008,19(1):82-89.
XU Y, LI J T, WANG B. A category resolve power-based feature selection method[J]. Journal of Software. 2008, 19(1): 82-89.

[10] ZHUO Y, ZHANG Q, GONG Z H. Research and implementation of network transmission situation awareness[A]. CSIE[C]. Los Angeles, USA, 2009. 210-214.

[11] 许俊明.图论及其应用[M]. 合肥: 中国科学技术大学出版社,2004.
XU J M. Graph Theory and Its Application[M]. Hefei: Publishing House of University of Science and Technology of China, 2004.

[12] FORD L R J, FULKERSON D R. A simple algorithm for finding maximal network flows and an application to the hitchcock problem[J].

Canada J Math, 1957, 9: 210-218.

[13] EDMONDS J, KARP R M. Theoretical improvements[J]. J Assoc Compute Math, 1972, 19: 248-264.

[14] KDD Cup 1999 Data[EB/OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[15] NLANR. http://pma.nlanr.net/Traces/Traces/long/auck/8/[EB/OL].

[16] LU Y. A Grid-based clustering algorithm for high-dimensional data streams[A]. Proc of the 1st International Conference on Advanced Data Mining and Applications[C]. LNCS, 2005. 824-831.

[17] Internet2. http://www.internet2.edu[EB/OL].

[18] BASS T. Defense-in-depth revisited: qualitative risk analysis methodology for complex network-centric operations[A]. Military Communications Conference (MILCOM), Communications for Network Centric Operations: Creating the Information Force[C]. 2001. 64-70.

[19] 陈秀真,郑庆华,管晓宏.层次化网络安全威胁态势量化评估方法[J]. 软件学报, 2006,17(4):885-897.
CHEN X Z, ZHENG Q H, GUAN X H. Quantitative hierarchical threat evaluation model for network security[J]. Journal of Software. 2006, 17(4): 885-897.

[20] 韦勇,连一峰,冯登国.基于信息融合的网络安全态势评估模型[J]. 计算机研究与发展,2009,46(3):353-362.
WEI Y, LIAN Y F, FENG G D. A network security situational awareness model based on information fusion[J]. Journal of Computer Research and Development, 2009, 46(3): 353-362.

[21] LAU S. The spinning cube of potential doom[EB/OL]. http://www.nersc.gov/nusers/security/TheSpinningCube.php, 2003.

[22] CONTI G, ABDULLAH K. Passive visual fingerprinting of network attack tools[A]. Proceedings of 2004 ACM Workshop on Visualization and Data Mining for Computer Security[C]. New York, USA, 2004. 45-54.

[23] KRASSER S, CONTI G, GRIZZARD J. Real-time and forensic network data analysis using animated and coordinated visualization[A]. Proceedings of the 2005 IEEE Workshop on Information Assurance[C]. 2005. 42-49.

[24] Carnegie Mellon's SEI. system for internet level knowledge (SILK)[EB/OL]. http://silktools.sourceforge.net, 2005.

[25] 杨雅辉,李小东.IP 网络性能指标体系的研究[J].通信学报,2002,23(11):1-7.
YANG Y H, LI X D. The study of a framework for IP network performance metrics[J]. Journal on Communications, 2002, 23(11): 2-7.

[26] 江勇,林闯,吴建平.网络传输控制的综合性能评价标准[J]. 计算机学报, 2002,25(8):870-887.
JIANG Y, LIN C, WU J P. Integrated performance evaluation criteria for network traffic control[J]. Chinese Journal Computers, 2002, 25(8): 869-877.

[27] 林闯,周文江,田立勤.IP 网络传输控制的性能评价标准研究[J]. 电子学报,2002,30(12A):1973-1977.
LIN C, ZHOU W J, TIAN L Q. Research on performance evaluation criteria for IP network traffic control[J]. Acta Electronica Sinica. 2002, 30(12A): 1973-1977.

[28] 张冬艳, 胡铭曾, 张宏莉. 基于测量的网络性能评价方法研究[J]. 通信学报, 2006,27(10):74-79.
ZHANG D Y, HU M Z, ZHANG H L. Study on network performance evaluation method based on measurement[J]. Journal on Communications, 2006, 27(10): 74-79.

[29] 蒋序平.网络性能综合评估方法 IEMoNP 的设计和实现[J].海军工程大学学报,2006,18(5):74-78.
JIANG X P. Design and realization of an integrated evaluation method of network performance[J]. Journal of Naval University of Engineering, 2006, 18(5): 74-78.

作者简介：

卓莹（1979-），女，江苏徐州人，国防科学技术大学博士生，主要研究方向为态势感知和网络管理。

龚春叶（1982-），男，湖南衡阳人，国防科学技术大学博士生，主要研究方向为并行计算和粒子输运。

龚正虎（1945-），男，湖南长沙人，国防科学技术大学教授、博士生导师，主要研究方向为计算机网络。