基于分布相似度迁移的关键路由设备检测

孟庆锴[1,2] 张 剡[2] 杨琬琪[2] 胡裕靖[2] 史颖欢[2] 潘红兵[1] 王 浩[3]
(南京大学电子科学与工程学院微电子设计研究所 南京 210046)[1]
(南京大学计算机软件新技术国家重点实验室 南京 210046)[2]
(华为技术有限公司南京研究所 南京 210012)[3]

摘 要 在基础设施网络(如电力网、互联网等)设施中,往往会出现关键节点,主要表现为节点流量大、在网络中位置关键等,其性能不稳定将制约网络部分区域的功能。因此从提高关键基础设施的性能和安全性的角度出发,针对关键基础设施的检测成为一个重要的研究课题。提出了一种新颖的基于分布相似度迁移的互联网关键路由设备的检测算法,其目的是自动地检测当前互联网线路中的关键路由设备。在真实环境中,不同线路中不同路由设备的行为特征由于若干客观因素(网络状态、路由设备性能等)导致其分布通常不相同。因此,所提方法主要基于路由之间的分布相似度迁移:首先在目标域(当前路由)中通过谱聚类方法自动判断出可疑的路由设备,然后通过提出的基于分布相似度迁移的分类器对上一步中检测出的可疑路由设备进行分类。在华为公司提供的真实数据集上进行的测试表明,所提方法能够有效发现线路中的关键路由设备,同时能够根据不同线路之间的分布相似度迁移来提高分类结果。

关键词 谱聚类,迁移学习,关键路由设备检测
中图法分类号 TP181 **文献标识码** A

Critical Routers Detection Based on Distribution Similarity Transfer

MENG Qing-kai[1,2] ZHANG Yan[2] YANG Wan-qi[2] HU Yu-jing[2] SHI Ying-huan[2] PAN Hong-bing[1] WANG Hao[3]
(Institute of VLSI Design, Nanjing University, Nanjing 210046, China)[1]
(State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China)[2]
(Nanjing Institute, Huawei Technologies, Nanjing 210012, China)[3]

Abstract Critical infrastructures, which usually have large flow and key position, are common in infrastructure networks (e. g. power transmission network, Internet). The performance and reliability of the critical infrastructures directly influence the local abilities of the whole networks. To improve the ability and security-level of infrastructure networks, we proposed a novel method for critical infrastructures detection, which is mainly based on distribution similarity transfer. The aim of the proposed method is to automatically detect the critical routers in the current route. In the real application, due to several factors (e. g. network status, performance of routers), the behaviors of different routers within different routes often belong to different distributions. Therefore, the proposed method models the problem as the distribution similarity transfer among different routes: First, the suspected routers are detected in the target domain (current route) by using spectral clustering; then, a newly proposed distribution similarity transfer classifier finally classifies the suspected routers obtained from the previous step. The proposed method was evaluated on the real dataset provided by Huawei Inst. The experimental results validate the proposed method can effectively detect the critical infrastructures. Meanwhile, it is demonstrated that the proposed method can successfully adopt the distribution similarity transfer to improve the classification results.

Keywords Spectral clustering, Transfer learning, Critical routers detection

1 引言

关键基础设施网络定义为一种独立的、大规模的、人造系统的网络,其基础设施能够协作和协同地产生一种持续性的流,包含基础物品流(如能源、数据和水)和服务流(如银行、医疗保健和运输)[1]。这些系统暴露于多种危害和威胁之下,例

到稿日期:2013-05-10 返修日期:2013-09-13 本文受国家自然科学基金项目(61035003,61175042,61021062),国家973项目(2009CB320702),江苏省973项目(BK2011005),教育部新世纪优秀人才支持计划(NCET-10-0476)资助。

孟庆锴(1989—),男,硕士生,主要研究领域为数据挖掘和云服务,E-mail:qingkaim@gmail.com;张 剡(1978—),男,讲师,主要研究领域为数据库设计和智能系统;杨琬琪(1988—),女,博士生,主要研究领域为机器视觉和数据挖掘;胡裕靖(1988—),男,博士生,主要研究领域为强化学习和智能Agent;史颖欢(1984—),男,博士,讲师,主要研究领域为医学图像分析与机器视觉;潘红兵(1971—),男,教授,硕士生导师,主要研究领域为多核处理器软件、CMOS传感器和自动化测量;王 浩(1975—),男,硕士,工程师,主要研究领域为IP网络和IP增值业务。

如负荷压力、恶意攻击等，因此需要对这种“复杂”系统的性能和可靠性进行评估。但实践中，这项评估已被证明为是一个不平凡的任务[2]。

互联网是由路由器等基础设施构建成的网络，因为需要转发大量报文，路由设备的负载能力对网络运行状态有着重要影响。在互联网线路中，存在一些路由设备，其主要表现为节点流量大、在网络中位置关键等。这些路由设备通常被认为是关键路由设备，从提高线路的安全性和性能的角度出发，如何自动检测线路中的关键路由设备是一个重要且具有挑战性的问题。

在该应用中，每条线路中有若干条不同的路由，每条路由是由不同的路由节点序列(至多30跳)组成，路由节点的数据为时间序列数据，表示当前路由节点相对前一跳在时间方向上的延时。关键路由设备检测的目标是针对当前路由判断是否存在关键路由设备，同时对关键路由设备进行自动检测，以提醒供应商、管理者注意。关键路由设备检测的主要难点有：(1)由于线路数量重多，且传输数据量大，人工检测方法的效率低且容易出错；(2)由于网络状态、路由设备性能等客观因素的影响，不同地区中不同线路数据特征往往不一样，无法通过简单的机器学习方法中的监督学习进行训练分类，部分数据特征如图1所示，横轴表示时间，纵轴表示相对前一跳的延时。

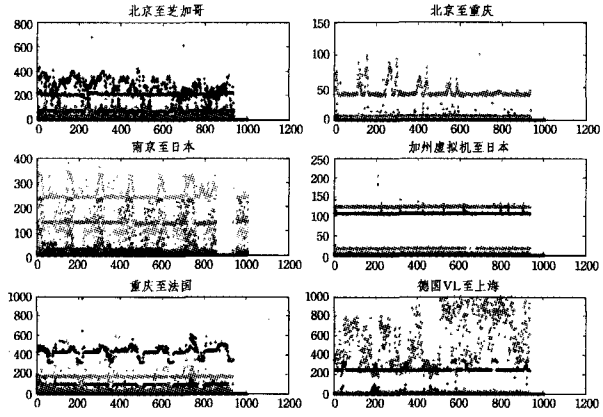

图1 部分路由特征

针对关键路由设备的自动检测可以归结为对时间序列的分析研究。近几年，研究者对时间序列分析做出了许多工作。典型在时间序列上的应用有分类[3,4]、聚类[5,6]、相似搜索[7]、预测[8]等；关于分类、聚类问题的算法有基于距离[5,6]、基于特征[9]、基于模型[10]和基于分割[11]等几种。其中，Wang[5]采取基于距离度量的聚类解决了时间序列的模式发现问题；Mörchen[9]基于时间序列特征抽取的问题，比较了DFT和DWT算法。

然而，由于传统的时间序列分析方法都假设训练线路和测试线路满足独立同分布假设，而在关键路由设备检测中，由于路由设备通常来自于不同地区的线路，存在着网络状态以及路由设备性能不相同等情况，因此我们需要针对该问题设计一种新的分析方法。本文提出的方法主要基于迁移学习[12]思想，将当前线路中的路由作为目标域，而其他线路中的路由作为源域。由于目标域中路由通常缺乏标记，我们的目标是通过源域的数据来帮助当前目标域的分析。方法的主要步骤是对每条路由中的路由节点进行基于相似度的聚类，得到疑似关键路由设备，但缺点是聚类数目难以确定；因为关键路由设备的延时序列的分布并不相同，所以对不同路由的路由设备延时序列直接通过相似度分析进行分类并不能达到预期的效果。本数据集具有层次结构，可以依据路由分布的相似度迁移关键路由设备的标记。据此，我们提出了基于分布相似度迁移的互联网关键路由设备检测方法，其结合路由分布以及路由节点的相似性来标记路由设备，判定疑似关键路由设备是否为关键路由设备，从而提高分类的准确度。

第2节介绍本文使用的测地线、谱聚类和最大均值差异算法以及使用的标记；第3节具体介绍基于分布相似度迁移的关键路由设备检测方法；第4节进行实验，对提出的算法进行性能测试并对实验结果进行分析；最后对本文工作做总结并展望下一步工作。

2 背景

在介绍本文提出的方法之前，将首先简要介绍测地线算法和谱聚类。

2.1 测地线算法

Peleg等人[13]将传统运输问题[14,15]引入计算机视觉领域，提出通过比较两幅灰度图像上像素点之间的灰度参数(pebble)的差异，来度量出匹配这两幅图的最小代价，可以将其作为两幅灰度图像之间的距离。Rubner等人[16]利用这个思想，提出了测地线算法(EarthMover's Distance，下称EMD)，它能够在图像检索中不同的特征空间上度量特征(signature)之间的距离，其可以理解为一种分布转化为另一种分布所需要的最小代价。EMD算法较为通用，具有较高的灵活性，在图像检索中有较好的性质，但EMD计算方法复杂，过大的特征空间会导致计算时间漫长，难以忍受。Pele等人在文献[17]中提出了$\widehat{\text{EMD}}$：

定义两个特征：

$$P=\{(\boldsymbol{p}_1,w_{p_1}),\cdots,(\boldsymbol{p}_m,w_{p_m})\} \tag{1}$$

$$Q=\{(\boldsymbol{q}_1,w_{q_1}),\cdots,(\boldsymbol{q}_n,w_{q_n})\} \tag{2}$$

其中，$\boldsymbol{p}_i$和$\boldsymbol{q}_i$为类标，w_{p_i}和w_{q_i}为其对应的权值。

定义矩阵$D=[d_{ij}]$为表示地面距离(ground distance)的矩阵，其中d_{ij}为类$\boldsymbol{p}_i$和$\boldsymbol{q}_i$之间的地面距离。定义流矩阵$F=[f_{ij}]$，其中f_{ij}为类$\boldsymbol{p}_i$和$\boldsymbol{q}_i$之间的流。$\widehat{\text{EMD}}$的目标是最小化式(3)：

$$\widehat{\text{EMD}}_\alpha=(\min_{\{f_{ij}\}}\sum_{i,j}f_{ij}d_{ij})+|\sum_i w_{p_i}-\sum_j w_{q_j}|\alpha\max_{i,j}d_{ij} \tag{3}$$

$$\text{s.t. } f_{ij}\geqslant 0,1\leqslant i\leqslant m,1\leqslant j\leqslant n$$

$$\sum_{j=1}^{n}f_{ij}\leqslant w_{p_i},1\leqslant i\leqslant m$$

$$\sum_{i=1}^{m}f_{ij}\leqslant w_{q_i},1\leqslant j\leqslant n$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}=\min(\sum_{i=1}^{m}w_{p_i},\sum_{j=1}^{n}w_{q_i})$$

$\widehat{\text{EMD}}$与EMD不同，它允许在非归一化的分布上使用。$\widehat{\text{EMD}}$能够加速近邻搜索、聚类，能够解决最大间隔(large margin)分类问题。

在此基础上，Pele和Werman[18]又提出了快速鲁棒的EMD算法，加速EMD或$\widehat{\text{EMD}}$的运算，具体方法是将$\widehat{\text{EMD}}$计

算中流的 N^2+N 条边通过引入一个转运顶点(transhipment vertex)缩减为 $N(K+3)$ 条边，如果 K 是一个常数，则算法时间复杂度将由 $\Theta(N^2)$ 降为 $O(N)$。本文使用的是快速鲁棒的 $\widehat{\text{EMD}}$ 算法。

2.2 谱聚类

谱聚类是建立在谱图理论上的一种聚类算法，根据谱图划分的思想，它将数据聚类问题转化成一个无向图的多路划分问题。谱聚类在很多领域都得到了成功的应用，包括机器视觉和 VLSI 设计[20,21]，近些年来在机器学习领域也得到了应用。

定义一个非空无向加权图 $G=(V,E)$，有限集合 $V=(v_1,\cdots,v_2)$ 表示图的顶点，集合 $E\subset V\times V$ 为顶点之间的边权重的集合。如果边 $(v_i,v_j)\in E$，那我们用 e_{ij} 表示这条边，很显然 E 是一个对称矩阵。将 E 看作是该无向图的邻接矩阵，它包含了聚类所需的所有信息。我们需要对 E 进行划分，从而得到不同的聚类，使得同一类中的点有较高的相似性，不同类之间的相似性较低。

对于图 G，定义它的邻接矩阵为 W，度(degree)矩阵为 $D(D_{ii}=\sum_{i\sim j}w_{ij})$，则该图的非归一化的图拉普拉斯(unnormalized graph Laplacian)矩阵定义为：

$$L=D-W \tag{4}$$

图拉普拉斯矩阵有以下性质：

1)对于任意向量 $(f_1,\cdots,f_n)'=f\in R^n$，以下等式一定成立：

$$f'Lf=\frac{1}{2}\sum_{i,j=1}^{n}w_{ij}(f_i-f_j)^2 \tag{5}$$

2)L 是一个对称的且半正定的矩阵。

3)L 的最小特征值为 0，对应的一个特征向量为 $\vec{1}$。

4)L 有 n 个非负的实数特征值，且

$0=\lambda_1\leqslant\lambda_2\leqslant\cdots\leqslant\lambda_n$

上述性质的相关证明见 von Luxburg[19] 的工作。

谱聚类的算法建立在以上理论基础之上，以 $k=2$ 的非归一化的算法为例，令 p 为 A 的划分指示向量：

$$f_j=\begin{cases}\sqrt{|A^c|/|A|}, & j\in A\\ -\sqrt{|A|/|A^c|}, & j\in A^c\end{cases} \tag{6}$$

需要最优化以下问题：

$$\begin{aligned}\min_{A\subset V}RatioCut(A,A^c):&=\min_{A\subset V}\frac{1}{2}\left(\frac{W(A,A^c)}{|A|}+\frac{W(A,A^c)}{|A^c|}\right)\\&=\min_{A\subset V}\frac{1}{2}\left(\frac{\sum_{i,j=1}^{n}w_{ij}(f_i-f_j)^2}{|V|}\right)\\&=\frac{\min f'Lf}{|V|}\end{aligned} \tag{7}$$

式中，$W(A,A^c)=\sum_{i\in A,j\in A^c}w_{ij}$，$A$ 和 A^c 为图的一个划分，$|V|$ 为 V 的顶点个数。

该最优化问题可以转化成为求解 L 的最小的两个特征值对应的特征向量所组成的矩阵的划分问题，以每一个特征向量为点的一个特征作为列向量，对矩阵的行进行 k 均值(k-Means)聚类即可得到最优化的一个图的一个划分。当 $k>2$ 时，可以推广为求解 L 的前 k 小的特征值对应的特征向量所组成的矩阵，聚类后即可得到所需划分。

3 基于分布相似度迁移的关键路由设备检测

3.1 对应于任务环境的记号

对应于任务环境的记号如表 1 所列。

表 1 对应于任务环境的记号

记号	记号名称	记号描述
V^s	服务器集合	所有服务器节点的集合 $V^s=\{x_1,\cdots,x_n\}$
x_i	服务器	标记为 x_i 的服务器节点
V^n	路由节点集合	数据中所有路由节点的集合 $V^n=\{n_1,\cdots,n_m\}$
n_i	路由节点	标记为 n_i 的各路由节点
E	相对延时集合	任意两两节点之间边的集合 $E=\{e_{v_1v_2},\cdots,e_{v_iv_j}\},(v_i,v_j\in V^s\cup V^n,i\neq j)$
$e_{v_iv_j}$	相对延时序列	节点 v_i 到节点 v_j 的延时时间序列
v_i	节点	表示网络中的一个服务器节点或路由节点
$R_{x_ix_j}$	路由集合	信源 x_i 到信宿 x_j 的路由集合 $R_{x_ix_j}=\{r_1^{x_ix_j},\cdots,r_k^{x_ix_j}\}$
$r_k^{x_ix_j}$	路由	信源 x_i 到信宿 x_j 的第 k 条路由 $r_k^{x_ix_j}=\{x_i\}\cup\{v_{t_1},\cdots,v_{t_{30}}\},(v_{t_p}\in V^s\cup V^n)$
$D_k^{x_ix_j}$	距离矩阵	描述路由 $r_k^{x_ix_j}$ 内部各节点之间的距离 $D_k^{x_ix_j}=[d_{v_mv_n}]v_m,v_n\in r_k^{x_ix_j}\backslash\{x_i\}$
$d_{v_mv_n}$	节点距离	路由 $r_k^{x_ix_j}$ 中节点 v_m 和 v_n 之间的标量距离
h(D)	归一化函数	使得 $0\leqslant d_{v_mv_n}\leqslant 1$
S	源域	由已标记的路由构成的知识迁移源域 $S=\{(r_k^{x_ix_j},D_k^{x_ix_j},W_k^{x_ix_j},y_k^{x_ix_j})\mid(i,j,k)\in G\}$
$y_k^{x_ix_j}$	标记向量	G 是节点对和路由组成的空间的一个划分 $r_k^{x_ix_j}$ 对应的标记向量
T	目标域	$T=\{r_k^{x_ix_j}\mid(i,j,k)\in G'\}$ G′是节点对和路由组成的空间的一个划分

3.2 互联网关键路由设备检测方法

源域 S 和目标域 T 中的路由分布不完全相同，难以直接用传统的分类方法直接分类，我们需要使用迁移的思想进行分类。

$r(r\in T)$ 的特点是关键路由设备与非关键路由设备的比例很小，通常情况下，关键路由设备的数量小于 6。对 r 进行聚类，配准后可以得到疑似关键路由设备 n。我们注意到数据存在层次结构，路由设备的标记是不能直接迁移的，需要结合路由分布的相似度和路由节点的相似度进行迁移，迁移公式如下：

$$H(n)=sgn\sum_{(i,j,k)\in G}f_{H_1}(r_k^{x_ix_j},r)g_{H_2}(r_k^{x_ix_j},n) \tag{8}$$

式中，f 在空间 H_1 上对 $r_k^{x_ix_j}$ 和 r 进行相似度度量，决定迁移的权重；g 在空间 H_2 上对 $r_k^{x_ix_j}$ 中的路由节点和路由节点 n 进行相似度度量，确定 n 的迁移的分类标记；sgn 表示参数的正负号，当参数大于 0 时返回 1，小于 0 时返回 −1。这样基于分布相似度的迁移可以有效避免负迁移。

3.3 算法描述

3.3.1 源域学习

我们的目标是发现延时相对较大、波动相对较大的路由节点。因此在度量节点间距离时，延时越大、波动越大的节点，距离相应也应该越大。EMD 算法符合这一特性，所以我们选择 EMD 算法进行节点间的相似度度量，节点 v_m 和 v_n 之间的距离如下：

$$d_{mn}=EMD(\boldsymbol{e}_{v_{m'}v_m},\boldsymbol{e}_{v_{n'}v_n})=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}^{g}f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}} \quad (9)$$

式中，$v_{m'}$ 和 $v_{n'}$ 分别为 v_m 和 v_n 在其所属路由中的前一个节点；d_{ij}^{g} 为地面距离。

据此，我们可以计算得到信源 x_i 到信宿 x_j 的第 k 个路由的距离矩阵 $D_k^{x_ix_j}$，其为 30×30 的方阵。

3.3.2 迁移

采用 EMD 算法计算 r 的距离矩阵 D^r，归一化得到 $W^r=[w_{v_mv_n}^r]=h(D^r)$。使用谱聚类对 W^r 进行聚类，关键路由设备的数量通常小于 6，配准后得到疑似关键路由设备 n，进一步通过迁移确认其是否为关键路由设备。聚类时最小化以下问题：

$$\min_{V\in r_k^{x_ix_j}} RatioCut(V_1,\cdots,V_k):=\min_{V_k\subset V}\frac{1}{k}\left(\frac{W(V_1,V_1^c)}{|V_1|}+\cdots+\frac{W(V_k,V_k^c)}{|V_k|}\right) \quad (10)$$

式中，$W(V_k,V_k^c)=\sum_{i\in V_k,j\in V_k^c}w_{ij}^r$，$|V_k|$ 为 V_k 的节点个数。

在迁移过程中，$r_k^{x_ix_j}$ 和 r 距离的度量采用最大均值差异算法（Maximum Mean Discrepancy，下称 MMD），该算法通过一个再生核希尔伯特空间（Reproducing Kernel Hilbert Space），能够方便地度量两个不同分布之间的差异（Borgwardt 等[22]）。

可以定义：

$$MMD(D_k^{x_ix_j},D^r)=\frac{1}{30\times29}f(D_k^{x_ix_j})+\frac{1}{30\times29}f(D^r)-\frac{2}{30\times30}f(DR) \quad (11)$$

式中，$f(D)=\sum_{m=1}^{30}\sum_{n=1}^{30}d_{mn}$，$DR=[d'_{v_mv_n}]$，$v_m\in r_k^{x_ix_j}$，$v_n\in r$。

则可以得到源域各路由与 r 的距离集合：

$$MMDS=\{MMD_k^{x_ix_j}\mid(i,j,k)\in G\} \quad (12)$$

式中，$MMD_k^{x_ix_j}$ 为 $r_k^{x_ix_j}$ 和 r 的距离。采用线性变换得到其对应相似度矩阵为：

$$MMDS'=\{\max(MMDS)-MMD_k^{x_ix_j}\mid(i,j,k)\in G\} \quad (13)$$

利用式(8)对路由标记进行迁移，迁移的权值依据式(13)相应相似度，迁移的类标为 $r_k^{x_ix_j}$ 中与 n 近邻的路由节点的类标，采用 KNN 方法决定。若关键路由设备的标记为 1，非关键路由设备的标记为−1，公式求和大于零时即认为该路由是关键路由设备。这样基于分布相似度的迁移可以有效避免负迁移，表 5 与表 6 也证明了我们的观点。

基于分布相似度迁移的关键路由设备检测算法的伪码描述如下：

算法伪码(r,S,Distance,Cluster,Similarity,Classifier)

Input：r：待分类的疑似关键路由设备

S：源域

Distance：度量路由节点之间相似度的算法(EMD)

Cluster：路由节点聚类算法(谱聚类)

Similarity：度量矩阵之间相似度的算法(MMD)

Classifier：分类算法(KNN)

$D^r\leftarrow d_{ij}=$Distance (n_i,n_j)

$W^r=h(D^r),(s_{ll}=0,l=\{1,\cdots,30\})$

C←Cluster (W^r)

配准后得到疑似关键路由设备集合$\{n_1,\cdots,n_k\}$

for h∈{1…k}

judgeLabel=0

while $(i,j,k)\in G$

judgeLabel+=f(Similarity $(r_k^{x_ix_j},n_h)$)·g(Classifier $(r_k^{x_ix_j},n_h)$)

end of while

Label$(n_i)\leftarrow$sgn(judgeLabel)

end of for

Output：Label(n)$\leftarrow\mathrm{sgn}\sum_{(i,j,k)\in G}f_{\mathbf{H}_1}(r_k^{x_ix_j},r)g_{\mathbf{H}_2}(r_k^{x_ix_j},n)$

4 实验

4.1 设置

为了验证所提方法在真实数据上的有效性，我们使用华为公司提供的真实数据。路由设备的延时、丢包信息相对容易获取，通过在世界各地租赁的 20 台服务器互相进行路由跟踪（如 windows 中的 tracert 命令），我们获取了格林威治时间 2013 年 1 月 18 日 13 时至 2013 年 1 月 24 日 12 时各服务器所跟踪的路由情况，并相应记录。主要记录有：

1）服务器到服务器之间每 5 分钟一次的路由跟踪获取到的路由节点；

2）从服务器到各路由节点的延时，频率为每秒一次；

3）发送包与接收包的时间戳和编号（用以计算丢包率）。

实验所用数据集如表 2 所列。其中数据集 1−5 为完全标记的数据集，数据集 6−10 都仅标记一条路由。

表 2 实验所用数据集

编号	数据集	路由数量	已标记路由
1	中国北京至中国重庆	12	12
2	中国北京至美国芝加哥	69	69
3	中国北京至德国法兰克福	50	50
4	中国南京至日本东京	10	10
5	美国加州虚拟机至日本东京	33	33
6	中国南京至中国重庆	8	1
7	中国重庆至法国巴黎	212	1
8	中国南京家庭至无锡虚拟机	162	1
9	德国法兰克福至法国巴黎	8	1
10	德国法兰克福至法国巴黎	36	1

除了源域和目标域同分布的情况外，本文还在源域和目标域分布不完全相同的情况下，对基于分布相似度迁移的关键路由设备检测算法进行了实验测试。同时，比较了在分布不完全相同的情况下，KNN 分类效果与基于分布相似度迁移的关键路由设备检测算法的差别。另外，比较了不同源域下，基于分布相似度迁移的关键路由设备检测算法的稳定性。实验使用的度量指标如下：

召回率 $R(recall)$=标记正确数/总关键设备数 (14)

精确度 $P(precision)$=标记正确数/标记数 (15)

$F1\ measure=2RP/(R+P)$ (16)

4.2 测试线路与训练线路同分布

在源域和目标域为相同分布的情况下，对算法进行测试。作为源域的路由个数为 2，剩余的作为目标域。对待标记路由中的路由节点进行谱聚类时，聚类数目为 3 个、4 个、5 个或 6 个，实验结果如表 3 所列；迁移后得到的结果如表 4 所列。

表3 谱聚类结果

已标记	待标记	待标记节点	聚类数	精确度	召回率	F1 Measure
1	1	300(10×30)	5	25%	100%	0.4
2	2	2010(67×30)	4	100%	76.90%	0.869
2	2	2010(67×30)	5	100%	88.89%	0.941
2	2	2010(67×30)	6	74.19%	100%	0.852
3	3	1440(48×30)	3	100%	96.89%	0.984
3	3	1440(48×30)	4	99.38%	96.76%	0.981
3	3	1400(48×30)	5	86.56%	100%	0.928
4	4	240(8×30)	5	75%	100%	0.857
5	5	930(31×30)	5	50%	100%	0.667

表4 分布相同时的基于分布相似度的迁移

已标记	待标记	待标记节点	聚类数	精确度	召回率	F1 Measure	执行时间
1	1	300	5	100%	100%	1	1208s
2	2	2010	4	100%	76.90%	0.869	6673s
2	2	2010	5	100%	88.89%	0.941	6848s
2	2	2010	6	99.67%	94.04%	0.968	7021s
3	3	1440	3	100%	96.89%	0.984	4995s
3	3	1440	4	100%	98.76%	0.994	5120s
3	3	1400	5	100%	98.76%	0.994	5320s
4	4	240	5	100%	100%	1	832s
5	5	930	5	100%	100%	1	3816s

可以看出单纯聚类的方法得到的分类结果具有不稳定性，准确率和召回率呈现反比关系，部分结果 F1 measure 较低。使用迁移的方法再次分类之后，准确率稳定，召回率受聚类结果的影响受到一定制约。通过提高聚类数目到一个合适的值，可以将召回率稳定在较高的水平。我们使用聚类的目的是减少算法运行时间，因此要注意聚类数目越多，得到的疑似关键路由设备越多，算法运行时间越长。实验显示算法执行时间相对较长，是因为 EMD 算法复杂度较高，单次两节点间距离度量需要 0.05s，而 MMD 中使用了大量 EMD 计算，导致总时间较长。

4.3 测试线路和训练线路分布不完全相同

进一步，在源域和目标域的分布不完全相同(但具有一定相似性)的情况下，对算法进行测试。如果不使用基于分布相似度的迁移，仅对源域的所有路由节点进行基于 EMD 的 KNN 分类，k 取 3 时，得到的实验结果如表 5 所列；使用我们的方法，得到结果如表 6 所列。

表5 分布不完全相同时的 KNN 分类

已标记	待标记	待标记节点	精确度	召回率	F1 Measure
1,3	5	990(33×30)	0%	0%	0
1,4	5	990(33×30)	0%	0%	0
2,4	5	990(33×30)	100%	100%	1
4,8	5	990(33×30)	0%	0%	0
7,10	5	990(33×30)	100%	100%	1
1,2,3,4	5	990(33×30)	100%	100%	1

表6 分布不完全相同的基于分布相似度的迁移

已标记	待标记	待标记节点	精确度	召回率	F1 Measure
1,3	5	990(33×30)	100%	100%	1
1,4	5	990(33×30)	100%	100%	1
2,4	5	990(33×30)	100%	100%	1
4,8	5	990(33×30)	100%	100%	1
7,10	5	990(33×30)	0%	0%	0
1,2,3,4	5	990(33×30)	100%	100%	1

数据集 2 的数据是中国(亚洲)到美国的路由，数据集 5 的数据是美国到日本(亚洲)的路由，其分布相似性较高，故 KNN 的分析结果也较好。尽管数据集 1、3、4 和 8 的数据与数据集 5 的相似性稍低，但是使用基于分布相似度的迁移的结果相对较好。数据集 7 和 10 的数据与数据集 5 相似性很低，即使基于分布相似度迁移，仍然不能进行良好的分类。

4.4 留一验证法结果

以上是在已知测试线路和训练线路样本分布的情况下做的分析。然而，在真实场景中，我们并不知道测试线路和训练线路样本分布的差异性。针对源域目标域相似度较低、不能很好迁移的问题，我们尝试通过构建一个较完备的源域，使其能够对待分类样本进行准确的标记。因此，我们对数据集 1—5做留一验证，实验结果如表 7 所列。

表7 不同源域上的迁移

已标记	待标记	待标记节点	精确度	召回率	F1 Measure
{1,…,10}\{1}	1	360(12×30)	100%	100%	1
{1,…,10}\{2}	2	2070(69×30)	98.26%	96.91%	0.976
{1,…,10}\{3}	3	1500(50×30)	94.41%	100%	0.971
{1,…,10}\{4}	4	300(10×30)	100%	76.47%	0.867
{1,…,10}\{5}	5	990(33×30)	100%	100%	1

在标记数据集 2 和 5 时，实验所用的方法欠佳，未能准确标记的关键路由设备有两种，即局部延时波动较大和延时波动较大但整体延时较小，这与我们实验中对关键路由设备的定义有关，因此需要进一步完善相似度的定义以发现此类关键路由设备。可以看出，在源域比较完备的情况下，算法表现出了非常好的稳定性。

4.5 k 值选取

实验中，算法使用 KNN 时 k 值选取为 1。这是因为每一条路由的关键设备数从 1～5 不等，k 值太大会影响投票的正确性。仍然采用留一验证，与表 7 不同的是，k 值取为 3，实验结果如表 8 所列。

表8 k=3 时不同源域上的迁移

已标记	待标记	待标记节点	精确度	召回率	F1 Measure
{1,…,10}\{1}	1	360(12×30)	0%	0%	0
{1,…,10}\{2}	2	2070(69×30)	100%	52.78%	0.691
{1,…,10}\{3}	3	1500(50×30)	93.46%	98.62%	0.960
{1,…,10}\{4}	4	300(10×30)	100%	57.69%	0.732
{1,…,10}\{5}	5	990(33×30)	100%	100%	1

结束语 本文提出了基于分布相似度迁移的关键路由设备检测算法。实验结果表明，基于分布相似度迁移的关键路由设备检测算法能够有效地利用迁移学习技术在数据复杂结构下进行高精度的分类。由于只需人工标记少量样本就可以进行剩余样本的标记，节省了大量的人力成本和时间成本。在进行节点聚类的过程中，使用节点间相似度作为邻接矩阵，通过变更相似度(距离)的定义，本文提出的基于分布相似度迁移的互联网关键路由设备检测算法可以适用于发现各种不同特征的关键路由设备，如局部异常等。另外，在目标域和源域分布不完全相同时，源域的组成将影响算法的稳定性，需要通过其他机器学习方法构造完备的源域来提升分类效果，这是需要进一步研究的内容。

参 考 文 献

[1] Ellis J, Fisher D, Longstaff T, et al. Report to the President's Commission on Critical Infrastructure Protection[R]. CARNEGIE-MELLON UNIV Pittsburgh PA Software Engineering INST, 1997

(下转第 40 页)

recommender systems research[J]. Expert Systems with Applications,2012,39(11):10059-10072

[2] 哈进兵,郑锐,甘利人. 一种基于加权关联规则的协同推荐算法[J]. 情报学报,2010,29(4):718-722

[3] 龙舜,蔡跳,林佳雄. 一个基于演化关联规则挖掘的个性化推荐模型[J]. 暨南大学学报,2012,33(3):264-267

[4] Thabtah F, Cowling P, Peng Y. MMAC: A new multi-class, multi-label associative classification approach[C]// ICDM 2004: Proceedings of the 4th IEEE International Conference on Data Mining. Brighton, UK, 2004:217-224

[5] 周欣,沙朝锋,朱扬勇. 兴趣度-关联规则的又一个阈值[J]. 计算机研究与发展,2000,37(5):627-633

[6] 肖波. 可信关联规则挖掘算法研究[D]. 北京:北京邮电大学,2009

[7] 李广原,杨炳儒,周如旗. 一种基于约束的关联规则挖掘算法[J]. 计算机科学,2012,39(1):244-247

[8] 杨红菊,梁吉业. 一种有效的关联规则的挖掘方法[J]. 计算机应用,2004,24(3):88-89

[9] 李杰,徐勇,王云峰. 面向个性化推荐的强关联规则挖掘[J]. 系统工程理论与实践,2009,29(8):133-151

[10] Liu Y Z, Jiang Y C, Liu Y C. CSMC: A combination strategy for multi-class classification based on multiple association rules[J]. Knowledge-Based Systems, 2008, 21(8):786-793

[11] Pang J F, Liang J Y. Evaluation of the results of multi-attribute group decision-making with linguistic information[J]. Omega, 2012, 40(3):294-301

[12] Jiang Y C, Shang J, Liu Y Z. Maximizing customer satisfaction through an online recommendation system: A novel associative classification model[J]. Decision Support System, 2010, 48(3): 470-479

[13] Cao F Y, Liang J Y, Li D Y, et al. A dissimilarity measure for the k-Modes clustering algorithm[J]. Knowledge-Based Systems, 2012, 26 (1):120-127

[14] 余力,刘鲁. 电子商务个性化推荐研究[J]. 计算机集成制造系统,2004,10(10):1306-1313

[15] Su J H, Wang B W, Hsiao C Y, et al. Personalized rough-set-based recommendation by integrating multiple contents and collaborative information[J]. Information Sciences, 2010, 180(1): 113-131

(上接第 31 页)

[2] Fang Y, Zio E. Unsupervised spectral clustering for hierarchical modelling and criticality analysis of complex networks[J]. Reliability Engineering & System Safety, 2013, 116:64-74

[3] Wei L, Keogh E, Xi X. SAXually explicit images: Finding unusual shapes[C]// Sixth International Conference on Data Mining, 2006. ICDM'06. IEEE, 2006:711-720

[4] Xi X, Keogh E, Shelton C, et al. Fast time series classification using numerosity reduction[C]// Proceedings of the 23rd international conference on machine learning. ACM, 2006:1033-1040

[5] Wang H, Wang W, Yang J, et al. Clustering by pattern similarity in large data sets[C]// Proceedings of the 2002 ACM SIGMOD international conference on management of data. ACM, 2002: 394-405

[6] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series[C]// Proceedings IEEE International Conference on Data Mining, 2001. ICDM 2001. IEEE, 2001:273-280

[7] Aβfalg J, Kriegel H P, Kröger P, et al. Similarity search on time series based on threshold queries[M]// Advances in Database Technology-EDBT 2006. Berlin Heidelberg: Springer, 2006:276-294

[8] Box G E P. Time Series Analysis: Forecasting and Control[M]. Wiley, 2008

[9] Mörchen F. Time series feature extraction for data mining using DWT and DFT[J]. 2003

[10] Ji X, Li-Ling J, Sun Z. Mining gene expression data using a novel approach based on hidden Markov models[J]. FEBS letters, 2003, 542(1):125-131

[11] Jiang J, Zhang Z, Wang H. A new segmentation algorithm to stock time series based on PIP approach[C]// International Conference on Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. IEEE, 2007:5609-5612

[12] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359

[13] Peleg S, Werman M, Rom H. A unified approach to the change of resolution: Space and gray-level[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(7):739-742

[14] Rachev S T. The Monge-Kantorovich mass transference problem and its stochastic applications[J]. Theory of Probability & Its Applications, 1985, 29(4):647-676

[15] Hitchcock F L. The distribution of a product from several sources to numerous localities[J]. J. Math. Phys., 1941, 20(2):224-230

[16] Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval[J]. International Journal of Computer Vision, 2000, 40(2):99-121

[17] Pele O, Werman M. A linear time histogram metric for improved sift matching[M]// Computer Vision-ECCV 2008. Berlin Heidelberg: Springer, 2008:495-508

[18] Pele O, Werman M. Fast and robust earth mover's distances[C]// 2009 IEEE 12th International Conference on Computer vision. IEEE, 2009:460-467

[19] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4):395-416

[20] Malik J, Belongie S, Leung T, et al. Contour and texture analysis for image segmentation[J]. International Journal of Computer Vision, 2001, 43(1):7-27

[21] Hagen L, Kahng A B. New spectral methods for ratio cut partitioning and clustering[J]. IEEE transactions on Computer-aided design of integrated circuits and systems, 1992, 11(9): 1074-1085

[22] Borgwardt K M, Gretton A, Rasch M J, et al. Integrating structured biological data by kernel maximum mean discrepancy[J]. Bioinformatics, 2006, 22(14):e49-e57