

基于深度学习的交通拥堵预测模型研究*

谭娟¹, 王胜春²

(1. 北京工商大学 商学院, 北京 100048; 2. 北京交通大学 交通数据分析与挖掘北京市重点实验室, 北京 100044)

摘要: 针对城市道路交通拥堵预警问题, 提出了一种基于深度学习的预测模型。通过归纳合并交通流参数、环境状态、时段等基础数据来构建交通流特征向量并确定四种预测状态。采用深度学习的自编码网络方法从无标签数据集中学习获取可表征数据深层特征的隐层参数并生成新特征集。应用 Softmax 回归对有标签的新特征集进行学习生成预测分类器。模型可对交通拥堵状况进行多态预测。通过仿真对比分析, 预测模型具有较省略特征学习的预测算法更好的预测性能, 平均预测精度可达 85%。

关键词: 交通拥堵; 预测模型; 深度学习; 自编码网络; Softmax 回归

中图分类号: TP181

文献标志码: A

文章编号: 1001-3695(2015)10-2951-04

doi: 10.3969/j.issn.1001-3695.2015.10.016

Research on prediction model for traffic congestion based on deep learning

Tan Juan¹, Wang Shengchun²

(1. Business School, Beijing Technology & Business University, Beijing 100048, China; 2. Beijing Key Laboratory of Traffic Data Analysis & Mining, Beijing Jiaotong University, Beijing 100044, China)

Abstract: In order to predict the traffic congestion in city, this paper proposed the prediction model based on the deep learning. The traffic data like flow parameters, environmental situation, times of day and so on were summarized and composed into the feature vector, and defined four kinds of traffic flow pattern. It learned the unlabelled traffic dataset by the auto-encoder network to build the hidden layer function and generate the new feature dataset, and used the softmax regression to classify the traffic pattern based on the labeled feature dataset learning. Demonstrated by the simulation, the prediction model has better performance than the model without the feature learning, the average prediction accuracy can reach 85%.

Key words: traffic congestion; prediction model; deep learning; auto-encoder network; Softmax regression

交通拥堵已经成为制约城市经济和社会发展的瓶颈,它直接造成城市的整体运转效率降低。在城市发展过程中的短效效应日益明显。国际诺贝尔经济学奖得主 Becker 曾经测算,全球每年因交通拥堵造成的损失占 GDP 的 2.5%。目前各国用来衡量交通拥堵的参数主要有拥堵时间、排队长度、车速等。中国公安部则对拥堵路段给出了定义:车辆在车行道上受阻且排队长度超过 1 km 的状态。它就像是一种慢性病,每天都在折磨着穿梭于城市道路上的人群,但应对起来却往往力不从心。

交通拥堵的治理应首在预防,即能根据道路的现有交通状态预测分析出短时间内的交通状态变化趋势,并对可能出现的拥堵现象进行预警;然后利用交通广播、微博等新媒体公众平台发出预警,疏导车辆合理选择行驶路线、加强秩序管理,以避免拥堵或缓解拥堵程度。因此,如何建立长效模型对交通拥堵进行及时预警是城市智能交通系统优化的研究热点。

交通拥堵预测模型研究就其学科性质而言属于信息科学的模式识别问题。目前国内外对于交通拥堵预测的诸多研究成果中也主要集中在该学科领域,主要有基于时间序列相关的预测分析^[1]、神经网络预测^[2]、贝叶斯网络预测^[3,4]以及多分类器组合预测^[5]等方法。这类方法的研究多见于理论研究,且由于其立足的基础理论在处理大数据上缺少鲁棒性,所以导

致模型普遍缺乏长效性和扩展能力。

深度学习^[6]是一种模拟人脑的多层感知结构来认识数据模式的学习算法。近年来作为数据挖掘的一个新兴领域,在处理图像、文本、语音等非结构化数据等方面体现出了极为卓越的性能,目前已逐渐被斯坦福大学、Google、百度研究院等权威机构作为 21 世纪数据挖掘和人工智能发展的战略方向。在交通应用领域, Schmidhuber 等人将该模型用于汽车智能驾驶的交通标志识别,使用深度学习的方式,获得了比人工识别还低的错误率^[7]。本文应用深度学习算法体系来处理交通流参数以及影响道路状态的环境数据、事关人流出行行为的时段分布等交通基础数据,构建了一个可对交通拥堵进行多态预测的长效模型。模型对于公路交通这一每天都在不断积累膨胀的数据样本,具有较好的自学习更新能力,可作为一种长效模型应用于智能交通管理系统。

1 交通状态特征数据遴选

现实中影响城市道路交通状态的因素是非常复杂的,既包含城市不同区域发展的差异化影响,这包括城市不同区域的人口密度差异、道路基础设施建设规格差异等,同时还受天气等环境因素的影响,如暴雨、暴雪等异常天气下必然更容易发生

收稿日期: 2014-07-16; 修回日期: 2014-08-28 基金项目: 北京市自然科学基金项目(9144022); 国家社科基金重点项目(13AGJ008); 国家教育部人文社科基金青年项目(12YJC630183)

作者简介: 谭娟(1983-),女,湖南邵阳人,讲师,博士,主要研究方向为交通运输规划、环境经济管理(tanjuan@btbu.edu.cn); 王胜春(1985-),男,山西朔州人,博士研究生,主要研究方向为机器学习、数据挖掘。

交通拥堵。另外就是一天不同时段人流出行的意愿对交通拥堵状态有直接影响。如上、下班的高峰时段容易发生道路拥挤,节假日高速通行免费、市政道路施工半封闭道路等因素势必造成道路拥堵。因此,本文从三方面来遴选用于表征交通拥堵影响因素的特征变量:a)交通流参数,这是最直观表述交通状态的参数,由交通部门的统计信息直接获得;b)环境因素,这部分数据需对定性描述进行定量转换处理,归并到模型的输入特征向量;c)时段因素,这里凸显早晚高峰人流变化、节假日、突发状况等因素。由于不同特征数据且量纲不同,所以必须根据经验知识对数据进行预处理,主要包括相关性分析、数据噪声的清理、输入数据的归一化。

1.1 交通流参数的选择

我国《道路交通阻塞度及评价方法(国标)》对于城市交通通行状况的描述主要从两个方面来定义评定指标,即交叉路口阻塞率和路段阻塞率。上述两个指标中交叉路口阻塞率和路段都是百分比参数。其中交叉路口阻塞率定义为车辆在交叉路口外车行道受阻排队长度超过500 m为阻塞,800 m为严重阻塞;路段阻塞率评定指标为长度超过2 000 m为阻塞,3 000 m为严重阻塞。阻塞率的定义分别如下:

$$\begin{aligned} \text{交叉路口阻塞率} &= \frac{\text{阻塞交叉路口数}}{\text{交叉路口数}} \times 100\% \\ \text{路段阻塞率} &= \frac{\text{阻塞路段数}}{\text{道路条数}} \times 100\% \end{aligned}$$

具体数据通过对关注区域的路段和交叉口进行固定周期测定或随机测定获得。固定周期测定时采用一组测量结果的若干个较高阻塞等级为评定结果。因此本文取用以上两个参数直接作为描述交通状态的两个变量,记为 CbR 、 RbR ,两个参数的取值均为 $[0, 1]$,因此无须再作归一化处理。

考虑到预测模型的时变性及交通流数据的稀疏性特征^[8],在交通流参数中引入高峰时段建成区主干道平均车速,用于评价道路的通畅程度,记为 Va 。该参数表示该观测路段的距离与路段上所有机动车辆的平均行车时间之比。然后根据不同的行车速度划定等级,如通畅: $Va \geq 30$;拥挤: $10 \leq Va < 30$;堵塞: $3 \leq Va < 10$;死锁: $3 \leq Va$,均以 km/h 作为计量单位。根据上文定义与车速相关的拥堵评价因子 B_v 。根据车速越低表明越拥堵,因此经分段归一化处理如式(1)所示。

$$B_v = \begin{cases} 0 & Va \geq 30 \\ 1 - \frac{Va - 3}{30 - 3} & 3 \leq Va < 30 \\ 1 & Va < 3 \end{cases} \quad (1)$$

交通流参数归结为以上三种指标。在模型的实际构建中。根据研究区域所关注的主干道、交叉路口数量, CbR 、 RbR 作为统计指标,在输入特征向量中为两个元素。拥堵评价因子 B_v 所占元素个数则根据主干道路来确定。因此,本文研究的模型为高维特征向量模型。

1.2 天气因素数据的处理

参考德克萨斯大学交通部对交通拥堵来源分类的数据研究,本文对于交通状态所涉及的气候因素的量化处理为五个等级,具体描述与天气预报的预警信号颜色关联。以 Tq 描述该指标,定义如下:

$$Tq = \begin{cases} 0.1 & \text{无预警} \\ 0.3 & \text{蓝色预警} \\ 0.5 & \text{黄色预警} \\ 0.7 & \text{橙色预警} \\ 0.9 & \text{红色预警} \end{cases} \quad (2)$$

1.3 时段因素的处理

时间因素 T 的处理,按照工作日和节假日分别处理。对于工作日,根据北京交管局车辆限行管理规定中对上下班高峰的定义:用0.9表示对应的早高峰7:00~9:00和晚高峰17:00~20:00时间;白天行车时间9:00~17:00用0.7表示;早高峰前6:00~7:00和晚高峰后20:00~22:00用0.3表示;其余时段用0.1表示。对于非工作日,日间时间8:00~20:00用0.5表示,其余时段用0.1表示。结合目前长假集中出行造成拥堵的现象,特征向量中引入假日 H 的影响。非假日时用0.1表示,周六、日时用0.3表示,其他法定的节假日按照休假的时间长短依次用0.5、0.7、0.9表示对交通的影响程度。

1.4 其他因素的影响

特殊事件 E 的影响,主要涉及市政工程、道路维修、交通事故、体育盛会等。道路无特殊事件时用0.1表示;当存有上述特殊事件时,可按照对交通状态的影响程度依次用0.3、0.5、0.7、0.9表示。该变量本文模型暂不作细致考虑,暂全设为0.1。

1.5 特征向量合成

综上所述,用 V 表示模型的输入特征向量,则向量可表示为 $V = [CbR, RbR, Tq, T, H, E, B_v(1), \dots, B_v(N)]$ 。其中 N 为模型所研究的区域内可测主干道路的数量,向量维度为 $N+6$ 。

城市道路交通拥堵状态的判别需要具体路况数据作为支撑,不同的城市会拥有自己独特的特点,按国际上通行的分类方法^[9]结合北京具体路况信息及交通拥挤指数以经验值确定四态作为预测输出 Z 的值域{死锁=0,堵塞=1,拥挤=2,通畅=3}。

2 深度学习模型设计

深度学习的概念源于人工神经网络的研究。含多层的多层感知器就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征,以发现数据的分布式特征表示。本文主要考虑到交通状态数据形式及本质特征,所以采用深度学习来构建预测模型。

每天都有大量的交通监控数据生成,并且拥堵现象出现的重复性高。这类真实存在的数据在其海量特性背后,必然隐含一定的共性特征。但在现象的梳理上,要对其进行归类整理和划分是一项繁重的工作。因此在挖掘技术的选择上,考虑只对交通监测数据作形式归纳,通过设计学习模型自动从无标签的监控数据中发现共性特征用于描述样本,然后通过仅对少量样本进行人工标志类别(交通状态判定输出)并作为输入到预测分类器训练的有标签数据集,最终形成一个可用于实际预测的模型。因此本文采用的是一种半监督学习模型。

2.1 基于自编码的深度学习模型

自编码(auto-encoder)是深度学习中一种快速学习模型,其基本原理利用了人工神经网络(ANN)的层次结构体系。在构建网络结构时,假设网络模型的输出与输入节点形式同构。本研究问题中,网络输入/输出节点形式基于交通监测参数构建的合成特征向量一致(参考2.5节)。网络模型的分布结构如图1所示。网络输入的每个节点对应特征向量 V 的一个元素,根据不同的预测需求中可测主干道路数量 N 确定输入节

点数量,并在输入节点中增广常数项节点 -1。网络输出节点也与特征向量 V 内元素形式逐一对应。定义用于自编码网络的学习训练集 $\mathcal{V} = \{v_1, v_2, \dots, v_n, n \in M\}$, 其中每个 v_i 即为一条交通数据采样样本, M 表示样本总规模。

本文采用的自编码网络为三层结构,包含输入层、单隐层、输出层,其结构模型如图 1 所示。与传统的神经网络直接通过学习然后用于预测问题不同,自编码网络仅关注隐层权重参数,并不执行分类操作。

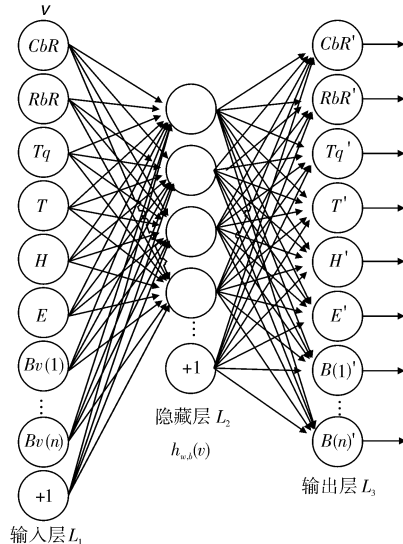


图 1 交通拥堵预测模型深度学习自编码网络结构

学习模型要求解的是满足逼近条件 $h_{w,b}(v) \approx v$ 隐层权重 w 的取值。深度学习理论认为该特征值是通过机器学习获取到的可表征样本模式库深层特征的新特征表达形式。用自编码网络学习的特征代替原始特征向量输入到分类器进行模式分类(即状态预测),可以大大提高分类精确度。在很多问题中甚至表现出超过目前最好分类算法的性能^[10]。

模型的求解采用的是通用的梯度下降法。求解过程实际为通过迭代逼近求取隐层权重 w 的过程。因输入特征向量都已作归一化处理,故采用 Sigmoid 函数作为隐层的变换核函数:

$$\begin{cases} z = w^T V + b \\ f(z) = \frac{1}{1 + \exp(-z)} \end{cases} \quad (3)$$

求得隐层权重 w 后,即可确定函数 $f(z)$ 为向量的特征变换核。自编码网络的学习功能所完成的是对交通参数特征数据样本深层特征的挖掘,并不提供对新样本的预测分类。接下来的工作是围绕学习获得的变换特征核设计可对交通状态进行预测的模式分类器。

2.2 基于 Softmax 的预测模型

在分类问题中,必须提供有标签的学习样本集供学习机作为参照。待分类器通过学习获得分类能力后才能对新输入的特征样本进行分类。本文采用 Softmax 模型建立预测分类器。

1) 构建训练集 设 x 是交通参数向量 v 经 $f(z)$ 函数变换后得到的新向量;根据先验知识对每个向量设定类别标签 $y^{(i)} \in \{1, 2, 3, 4\}$, 分类代表预测模型要输出的四种状态。由此得到有标签学习向量集 $L = \{(x^1, y^1), \dots, (x^m, y^m)\}$ 。

2) 求解预测分类器 Softmax 模型是 logistic 模型在多分类问题上的推广。对于给定的训练样本集合输入 L ,可考虑采用一个假设函数针对每个类 j 估算概率值 $p = (y = j | x)$ 。于是

定义假设函数 $h_\theta(x)$ 形式如下:

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1 | x^{(i)}; \theta) \\ p(y^{(i)} = 2 | x^{(i)}; \theta) \\ p(y^{(i)} = 3 | x^{(i)}; \theta) \\ p(y^{(i)} = 4 | x^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^4 e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix}$$

其中: $\theta_1, \theta_2, \dots, \theta_k$ 为要求取的模型参数。同样地,通过定义代价函数:

$$J(\theta) = -\frac{1}{m} \left[\sum_{i=1}^m \sum_{j=1}^k 1\{y^i = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^k e^{\theta_l^T x^{(i)}}} \right]$$

并求取使代价函数最小的模型参数。通过迭代确定参数 θ_k , 即可获得最终的模式预测分类器。

3 仿真和验证

模型的仿真和验证工作基于北京市包含 42 条主干道路的某局部区域的监控数据开展。样本收集工作依据官方提供的对该区域交通历史监控数据进行梳理,并将监控数据与天气、时段等其他特征因素合并,根据主干道路 $N = 42$ 确定向量维数为 48,向量特征结构参照 2.5 节定义。根据预测模型的构建需求,需分别整理两类学习集,即用于确定 $f(z)$ 的特征学习集和用于确定 Softmax 分类模型参数的分类学习集。两个学习集的整理流程如下所示: a) 特征学习集,设定特征学习集样本规模为 50 000,为保证特征学习的普适性,采用随机抽取方式从总样本集中选择; b) 分类学习集,设定分类学习集样本规模为 10 000,分类学习集需要为集合中每个样本特征向量要设置类别标签,类别标签与 2.5 节定义一致。其中 5 000 组样本用于训练 Softmax 分类器,另外 5 000 组用于验证模型性能。学习集整理完成后,仿真和验证工作按照图 2 所示流程开展。

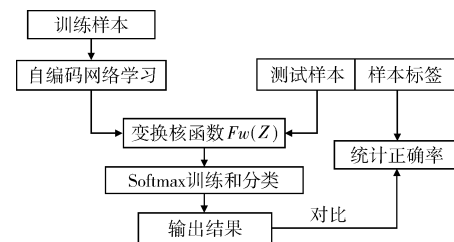


图 2 系统仿真验证操作流程

仿真工作在 Windows 7 64 位系统 Intel-i7 CPU、32 GB 内存的高配置计算机上进行,编码实现在 MATLAB 2012R 平台下完成。自编码网络采用上文论述的三层结构(输入、隐层、输出),输入层节点数为 49,隐层节点数设为输入节点数的 1/2,输出节点数为 48。仿真工作首先验证学习模型在学习集规模不断扩展情况下的进化能力。实验过程如下:

a) 设定学习集初始规模为 5 000,并以 5 000 为步长递增;直到前期数据整理的学习集规模上限为 50 000。

b) 在不同学习集规模上采用本文方法进行特征学习,特征学习的迭代上限为 400,并基于学习后特征对 Softmax 分类器进行训练,然后在测试集验证分类精度。为方便性能比较,同时采用支持向量机(SVM-RBF 方法)以及 Softmax 分类器在原始数据向量上进行学习并验证精度(即学习样本集未经自编码网络进行特征学习)。10 组不同学习集的预测分类精度如图 3 所示。

从图 3 中可以看出,在学习样本规模较小时,基于原始特征向量的 SVM-RBF 预测分类精度较高。本文方法的预测精度

最差,准确率不到50%。但随着输入的交通数据训练样本的增加,本文方法预测的准确率却是随着学习集规模增长而逐渐上升的,最终维持在85%左右;而未采用自编码学习特征的SVM预测随着学习集规模的增加,其预测准确率逐渐衰减,但维持在70%左右。同样对比了省略特征学习过程的Softmax直接分类器,其性能随着学习集规模的增加而不显著。

通过对比发现,本文提出的预测模型凸显了对交通监测数据的二次特征学习,其性能较无特征学习的直接预测分类具有更好的增量学习特性,这个优点十分适合于对交通流这类数据进行挖掘处理。

再对方法的泛化性能进行仿真验证。采用在完整学习集上训练完成的分类器在新数据集进行测试。新集包含1 000组样本。每组样本内信息容量为10 000条。其中第1~300组对应的是早高峰时段采样数据;301~650组对应的是非高峰时段采样数据;651~1 000组对应的是晚高峰时段采样数据。计算的预测准确率的粉笔曲线如图4所示。分析曲线的规律,数据处理的正确率总体呈均值分布,仅个别样本组预测结果存在异常波动,这表明系统具有较好的预测稳定性。

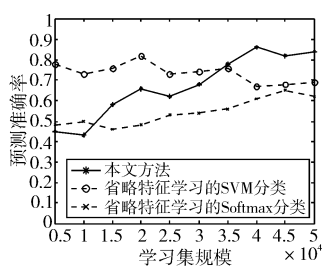


图3 本文方法精度与省略特征学习性能对比

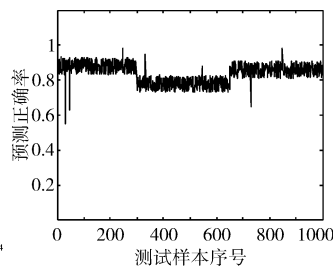


图4 交通拥堵预测模型准确率统计

从三组不同属性的样本特性来看,对于早晚高峰期存在拥堵的情况,预测准确率统计均值分别维持在86.7%和84.2%的水平;对于非高峰期的数据样本,则维持在78.9%的水平。预测系统在处理真实拥堵数据时的性能要优于非拥堵状态。分析其原因,最大可能性是学习集内可能包含了更多的拥堵时段数据,因此系统在处理真实拥堵数据时的准确率更高。由此可见,应用自编码网络的特征学习在对交通数据认知层次确实具有可挖掘的深度辨识能力。

4 结束语

基于深度学习的认知机理,构建自编码网络挖掘交通拥堵数据流隐含的深层特征,并结合Softmax回归方法构建分类器,设计并实现了一个适用于交通特征数据挖掘和交通拥堵预测的模型。经仿真验证,模型具有较好的增量学习特性,其对交通拥堵高峰期的预测精度约为85%。

研究工作的远景设想旨在能构建一个适用于大型城市交通流数据高鲁棒性的预测系统。本文仅是该项工作的数据挖掘及模式分类核心模块的研究工作之一,其更重要的研究价值在于基于其学习机理可在并行计算网络上构建更高层级的深度学习网络,最终研究目标是要实现可对千万级别公路交通数据集进行特征挖掘和预测的模型。

参考文献:

- [1] Lo H K, Luo X W, Siu B W Y. Degradable transport network: travel time budget of travelers with heterogeneous risk aversion[J]. Transportation Research Part B: Methodological, 2006, 40(9): 792-806.
- [2] 沈小军, 陈峻, 王晨. 基于LVQ神经网络的交通拥堵预测研究[J]. 交通运输系统工程与信息学报, 2009, 7(3): 97-102.
- [3] 陆宇, 岳昆, 刘惟一. 一种基于贝叶斯网络的交通拥堵预测方法[J]. 云南大学学报: 自然科学版, 2010, 32(S1): 355-363.
- [4] 王建, 邓卫, 赵金宝. 基于贝叶斯网络多方法组合的短时交通流量预测[J]. 交通运输系统工程与信息, 2011, 11(4): 147-153.
- [5] 李春英, 汤志康, 曹元大. 多分类器组合的交通拥堵预测模型研究[J]. 计算机工程与设计, 2010, 31(23): 5088-5091.
- [6] Hinton G E, Osindero S, Teh Y. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.
- [7] Ciresan D C, Meier U, Masci J. Multi-column deep neural network for traffic sign classification[J]. Neural Networks, 2012, 32(8): 333-338.
- [8] 祁伟, 李晔, 汪作新. 季节性ARIMA模型在稀疏交通流下的预测方法[J]. 公路交通科技, 2014, 31(4): 130-135.
- [9] 李志恒, 孙东, 靳雪翔, 等. 基于模式的城市交通状态分类与性质研究[J]. 交通运输系统工程与信息, 2008, 8(5): 83-87.
- [10] Shin H C, Matthew R O, David J C. Stacked auto-encoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data[J]. Pattern Analysis and Machine Intelligence, 2013, 35(8): 1930-1943.
- [8] Jiang Chunxiao, Chen Yan, Liu K J R. Evolutionary dynamics of information diffusion over social networks[J]. Signal Processing, 2013, 62(17): 4573-4586.
- [9] Remy C, Pervin N, Toriumi F, et al. Information diffusion on Twitter: everyone has its chance, but all chances are not equal[C]//Proc of International Conference on Signal-Image Technology & Internet-Based Systems. [S.l.]: IEEE Press, 2013: 483-490.
- [10] Jiang Chunxiao, Chen Yan, Liu K J R. Modeling information diffusion dynamics over social networks[C]//Proc of International Conference on Acoustics, Speech and Signal. [S.l.]: IEEE Press, 2014: 1095-1099.
- [11] Hajibagheri A, Hamzeh A, Sukthankar G. Modeling information diffusion and community membership using stochastic optimization[C]//Proc of International Conference on Advances in Social Networks Analysis and Mining. [S.l.]: IEEE Press, 2013: 175-182.
- [12] Niu Jianwei, Huang Shaluo, Shu Lei, et al. Activities information diffusion in Chinese largest recommendation social network: patterns and generative model[C]//Proc of Global Communications Conference. [S.l.]: IEEE Press, 2013: 3083-3088.

(上接第2950页)

- [3] Myers S A, Zhu Chenguang, Leskovec J. Information diffusion and external influence in networks[C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2012: 33-41.
- [4] Budak C, Agrawal D, El Abbadi A. Diffusion of information in social networks: is it all local? [C]//Proc of the 12th International Conference on Data Mining. [S.l.]: IEEE Press, 2012: 121-130.
- [5] Ghosh R, Lerman K. A framework for quantitative analysis of cascades on networks[C]//Proc of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2011: 665-674.
- [6] Saito K, Kimura M, Ohara K, et al. Efficient discovery of influential nodes for SIS models in social networks[J]. Knowledge and information systems, 2012, 30(3): 613-635.
- [7] Wang Feng, Wang Haiyang, Xu Kuai, et al. Characterizing information diffusion in online social networks with linear diffusive model[C]//Proc of the 33rd International Conference on Distributed Computing Systems. [S.l.]: IEEE Press, 2013: 307-316.