一种基于随机森林的LBS用户社会关系判断方法

马春来 单 洪 马 涛 顾正海
(电子工程学院 合肥 230037)

摘 要 根据LBS用户位置信息对用户之间是否存在社会关系进行判断,是基于位置大数据的情报挖掘领域中的一个新兴问题,可为群体发现及社团划分提供信息支撑。以时空共现理论为依据,将时空共现区特征归纳为4类,提出了一种基于随机森林的用户社会关系判断方法。该方法包括特征选择和训练分类环节。首先,针对特征空间存在不相关和冗余特征而影响判断性能的问题,提出一种基于Fisher准则和χ^2检验的特征选择算法,对无关、冗余特征进行剔除;然后采用随机森林进行分类判断,克服了现有方法训练速度慢、容易过拟合的问题。以LBSN用户Check-in数据为例进行的实验结果表明,该方法能够以较低的计算代价和较高的准确率实现社会关系的判断。

关键词 基于位置的服务,时空共现,随机森林,社会关系推断
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.040

Random Forests Based Method for Inferring Social Ties of LBS Users

MA Chun-lai SHAN Hong MA Tao GU Zheng-hai
(Electronic Engineering Institute, Hefei 230037, China)

Abstract Inferring social ties from the location information of LBS users, which can provide more information for group discovery and community detection, is now becoming a new problem in intelligence mining from location big data. Based on the theory of co-occurrences, the features of co-occurrences region were divided into four categories, and a new method based on random forests for social ties inferring was proposed in this paper. The method consists of feature selection phase and classification phase. Firstly, for the problem that uncorrelatedand redundant features will affect the accuracy of result, an algorithm based on Fisher criterion and χ^2 test was proposed to remove the uncorrelated and redundant features. Secondly, random forests was applied in the classification to overcome the problem of existing method that training phase is slow and the model is easily over-fitting. Check-in data of LBSN users is chosen as test data in experiment, the results indicate the feasibility and effectiveness of the method.

Keywords LBS, Spatio-temporal co-occurrences, Random forests, Social ties inferring

近年来,移动通信、互联网、GNSS(Global Navigation Satellite System)等技术的普及极大地促进了LBS的发展[1],使得用户的历史轨迹信息达到了较大规模。这些信息代表了某种事件或者活动,而这些事件或活动隐含着人类的行为模式和生活习惯[2,3]。因此,合理地使用时空数据挖掘技术可对该类数据进行有效的知识发现及价值提取[4]。研究表明,人类的社会关系与其轨迹信息具有一定联系。这就意味着,在获取用户历史轨迹的前提下可对两个目标用户之间是否存在社会关系进行推断,在此基础上,可进一步进行群体识别及社团发现[5-7],从而在舆情管控、打击恐怖主义及其他犯罪组织、维护社会稳定等方面提供情报保障[7-9]。

社会关系推断主要包括3个方面的应用:1)对两用户是否存在社会关系进行判断;2)对社会关系类型进行推断[10];3)对用户的社会关系结构进行推断[5]。目前国内研究刚处于起步阶段,国外研究主要集中于第一方面的研究。社会关系的判断主要通过3种方法:1)通过用户到访时空共现(Spatio-Temporal Co-occurrences)区的频次建立概率模型进行判断[11];2)通过用户之间轨迹的相似性,对社会关系进行判断[12];3)通过机器学习方法[13,14],对用户多种时空特征(到访频次、到访时间、持续强度、位置属性等)进行训练学习并分类判断。

其中,第3种方法综合了多种信息作为参考,具有较高的准确率。但由于特征较多、维度较高,使得传统的机器学习方法难以处理此问题。鉴于此,提出了一种基于随机森林的LBS用户社会关系判断方法,该方法首先对用户的时空特征进行合理的人工分类;然后提出一种基于Fisher比和Person χ^2检验的过滤型特征选择方法,分别剔除不相关特征和冗余特征;最后采用随机森林方法进行分类并判断。

1 时空共现的概念与总体思路

1.1 时空共现的概念

结合"Co-occurrences"及"Co-location",本文给出"时空共现区"的定义:如图1所示,区域S_i为$l\times l$大小的空间区域,T_i为起始时间t_s,持续时间为T的时间段。假设两个用户

到稿日期:2015-10-09 返修日期:2016-01-25 本文受国防重点实验室基金资助。
马春来(1989-),男,博士生,主要研究方向为机器学习、基于位置大数据的情报挖掘,E-mail:eviiive@163.com(通信作者);单 洪(1965-),男,教授,博士生导师,主要研究方向为战场无线网络、大数据情报挖掘;顾正海 男,硕士,讲师,主要研究方向为战场无线网络、机器学习。

u_1，u_2 在同一时间段 T_i 均出现在同一区域 S_i，则 T_i 及 S_i 所约束下的区域被称为用户 u_1，u_2 的时空共现区。

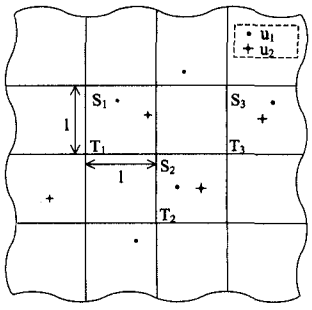


图 1　时空共现区

图 2 给出了不同时空共现区大小与样本数的关系，可以看出，随着时空共现区的增加，样本总数迅速增加，但这并不意味着社会关系的判断更加准确。研究表明，在合理的参数设置下，时空共现事件与用户是否存在社会关系密切相关[13]。

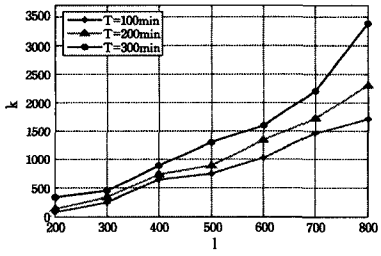


图 2　时空共现区大小与样本数的关系

1.2　总体思路

所提方法的总体思路可分为数据提取、特征选择、训练、分类 4 个环节，如图 3 所示。

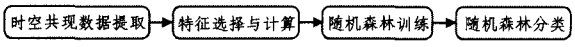


图 3　社会关系判断的总体思路

1）设定区域大小 l 及时间间隔长度 T，搜索该约束下目标用户的时空共现区；2）采用 Filter 型特征选择方法，对已经分类的特征进行选择，根据选择好的特征，计算时空共现区内每一对用户的特征取值，生成 k 条实验数据；3）采用随机森林算法对训练数据进行学习，构建 k 个 CART 树；4）根据 k 个 CART 树的投票结果对社会关系进行分类判断。在实验验证部分中，由于时空共现数据是从用户的历史轨迹信息中提取的，考虑到带有 LBS 用户社会关系标签的公开数据较少，采用 LBSN 用户的 Check-in 数据，并将用户间的关注关系作为标签数据，以验证所提方法的可行性和有效性。下面将着重介绍特征的分类、选择与随机森林算法。

2　特征选择与随机森林算法

2.1　特征分类

Justin Cranshaw 等人虽然归纳了多种特征，但这些特征对实验数据要求较高，由于 LBSN 用户更新位置属于偶发更新[15]，因此数据属性呈稀疏特性。鉴于此，从时域、空域、频域、用户等方面重新定义总体属性、用户活跃性、位置多样性、位置特殊性等 4 方面的特征。

2.1.1　总体属性

用户在时空共现区签到的总体属性主要描述用户时空数据的整体概貌，包括用户总数、区域总数、到访时长等信息，如表 1 所列。

表 1　总体属性

类别	变量	意义	数量
总体属性	N_U	用户总数	1
	N_L	区域总数	1
	$N_U^{L_i}$	所有用户到访时空共现区 L_i 的数目	1
	$N_{u_i}^{L}$	用户 u_i 出现在所有时空共现区 L 的总数目	2
	$T_{u_m}^{L}$，$T_{u_n}^{L}$	用户 u_m，u_n 出现在所有时空共现区 L 的总时间	2

2.1.2　用户活跃性

用户在时空共现区的活跃程度，可用两用户到访时空共现区的总数目，到访该位置的用户总数量，在不同时间段两用户到访该位置的次数，两用户到访该位置的日期、时间及间隔等参数来表示，如表 2 所列。

表 2　用户活跃性

类别	变量	意义	数量
用户活跃性	N_{u_m,u_n}^{L}，$N_{u_m,u_n}^{L,E}$，$N_{u_m,u_n}^{L,W}$	用户 u_m，u_n 出现在所有时空共现区 L 的总数目（晚上、周末时间段总数目）	3
	$N_{u_m,u_n}^{L_i}$，$N_{u_m,u_n}^{L_i,E}$，$N_{u_m,u_n}^{L_i,W}$	用户 u_m，u_n 出现在时空共现区 L_i 的总数目（晚上、周末时间段总数目）	3
	$T_{u_m}^{L_i}$，$T_{u_n}^{L_i}$	用户 u_m，u_n 出现在时空共现区 L_i 的总时间	2
	$N_{u_m,u_n}^{L_i,M}$，$N_{u_m,u_n}^{L_i,W}$，$N_{u_m,u_n}^{L_i,H}$	$N_{u_m,u_n}^{L_i,M}$ 表示一个月中的第几天，$N_{u_m,u_n}^{L_i,W}$ 表示一周中的第几天，$N_{u_m,u_n}^{L_i,H}$ 表示一天中的第几小时	3
	$maxD_{u_m,u_n}^{L_i}$，$medD_{u_m,u_n}^{L_i}$，$minD_{u_m,u_n}^{L_i}$，$avrD_{u_m,u_n}^{L_i}$，$varD_{u_m,u_n}^{L_i}$	用户 u_m，u_n 出现在时空共现区 L_i 的直线距离（最大值、中值、最小值、平均值、方差）	5
	$maxF_{u_m,u_n}^{L_i}$，$medF_{u_m,u_n}^{L_i}$，$minF_{u_m,u_n}^{L_i}$，$avrF_{u_m,u_n}^{L_i}$，$varF_{u_m,u_n}^{L_i}$	用户 u_m，u_n 出现在时空共现区 L_i 的频率（最大值、中值、最小值、平均值、方差）	5

2.1.3　位置多样性

位置多样性采用空间位置熵[16] $E_{u_j}^{L_i}$ 进行度量，用于评价用户 u_j 在区域 L_i 的可预测程度，定义如式(1)所示：

$$E_{u_j}^{L_i}=-\sum_{j=1}^{N_{u_j}^{L_i}}P_{u_j}^{L_i}\log P_{u_j}^{L_i} \tag{1}$$

$$P_{u_j}^{L_i}=\frac{N_{u_j}^{L_i}}{N_U^{L_i}} \tag{2}$$

其中，$N_{u_j}^{L_i}$ 为用户 u_j 到访区域 L_i 的次数，$N_U^{L_i}$ 为到访位置 L_i 的总数量，$P_{u_j}^{L_i}$ 为用户 u_j 到访区域 L_i 的概率。

为进一步说明位置熵的意义，采用 Foursquare 用户的 Check-in 数据进行实验，划分时空共现区并根据式(1)计算各点位置熵，如图 4 所示，根据颜色色调及位置点的密集程度表示熵值大小。由图 4 可看出，公共场所一般具有较高的熵值，这是由于公共场所到访的用户数目较多，用户随机性更强，可预测性就愈差。而个人住宅区用户的熵值较低，这是由于相比于公共区域，到访个人住宅区的用户数较少。

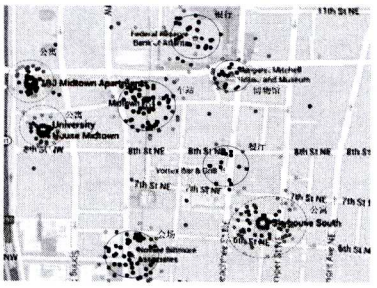


图 4　Foursquare 用户位置熵分布

上述实验表明，位置熵值越大，信息增益越大。通常公共场所（商场、饭店、娱乐场所等）的位置熵要比个人场所（住宅、工作单位）的熵值略大。这是时空共现区属性的主要特征。

表 3　位置多样性

类别	变量	意义	数量
位置多样性	$\max N_{u_i}^{L_i}$，$\mathrm{med}N_{u_i}^{L_i}$，$\min N_{u_i}^{L_i}$，$\mathrm{avr}N_{u_i}^{L_i}$，$\mathrm{var}N_{u_i}^{L_i}$	用户 u_i 到访时空共现区 L_i 的频率（最大值、中值、最小值、平均值、方差）	10
	$\max E_{u_j}^{L_i}$，$\mathrm{med}E_{u_j}^{L_i}$，$\min E_{u_j}^{L_i}$，$\mathrm{avr}E_{u_j}^{L_i}$，$\mathrm{var}E_{u_j}^{L_i}$	时空共现区 L_i 对用户 u_j 的位置熵（最大值、中值、最小值、平均值、方差）	5

2.1.4　位置特殊性

所谓位置特殊性，是指某一位置对用户的意义和关联程度，可通过频繁程度、分布范围及重要程度来表示。为了评估时空共现区 L_i 对两用户 u_1，u_2 的重要程度，根据 TF-IDF 理论引入词频 TF 参数，如式(3)所示。

$$TF_{u_1,u_2}^{L_i}=\frac{N_{u_1,u_2}^{L_i}}{N_U^{L_i}} \tag{3}$$

其中，$N_U^{L_i}$ 为时空共现区到访的所有用户数量，如式(4)所示。

$$N_U^{L_i}=\sum_{j=1}^{N_U}N_{uj}^{L_i} \tag{4}$$

用户 u_m，u_n 到访时空共现区 L_i 的次数占两用户到访区域次数总和之比为 $R_{u_m,u_n}^{L_i}$，用以表征两用户到访 L_i 的频繁程度，如式(5)所示。

$$R_{u_m,u_n}^{L_i}=\frac{\sum_{i=1}^{N_l}N_{u_1,u_2}^{L_i}}{\sum_{i=1}^{N_l}N_{u_1}^{L_i}+\sum_{i=1}^{N_l}N_{u_2}^{L_i}} \tag{5}$$

用户 u_m，u_n 到访不同时空共现区数目占所有用户到访区域数总和之比为 R_{u_m,u_n}^{P}，用以表征两用户的分布范围，如式(6)所示。

$$R_{u_m,u_n}^{P}=\frac{N_{u_i}^{P}}{\sum_{i=1}^{N_U}N_{u_i}^{P}} \tag{6}$$

表 4 列出了位置特殊性参数详细说明。

表 4　位置特殊性

类别	特征	变量	意义	数量
位置特殊性	频繁程度	$R_{u_m,u_n}^{L_i}$	表征用户 u_m，u_n 到访时空共现区的频繁程度	1
	分布范围	R_{u_m,u_n}^{P}	表征用户 u_m，u_n 的分布范围	1
	重要程度	$\max TF_{u_1,u_2}^{L_i}$，$\mathrm{med}TF_{u_1,u_2}^{L_i}$，$\min TF_{u_1,u_2}^{L_i}$，$\mathrm{avr}TF_{u_1,u_2}^{L_i}$，$\mathrm{var}TF_{u_1,u_2}^{L_i}$	表征时空共现区 L_i 对两用户 u_1，u_2 的重要程度	5

2.2　特征选择

在以上 4 类共计 50 个特征中，$T_{u_m}^{L}$，$T_{u_n}^{L}$，$T_{u_m}^{L_i}$，$T_{u_n}^{L_i}$ 4 个特征在偶发更新的 Foursquare 数据集中无法统计和计算，故本文拟确定 46 个特征作为判别社会关系的输入。然而，在该 46 个特征中存在不相关特征和冗余特征，从而影响了判断准确率。鉴于此，提出一种基于 Fisher 准则和 χ^2 检验的特征选择方法。

该方法首先采用 Fisher 比对特征的重要性进行度量，对低于阈值 α 的特征进行剔除，然后以采用 Pearson χ^2 检验计算剩余候选特征集中各特征之间的相关性，最后将 χ^2 值高于阈值 β 的特征进行剔除。

(1)重要性度量

现假设数据集共有 k 个样本，且分属于 C 个类别，第 w 类样本的个数为 k_w，第 w 类中第 r 维特征的均值为 μ_{wr}，全部样本中第 r 维特征的均值为 μ_r。

Fisher 比可表示为式(7)：

$$f_r=\frac{v_B^r}{v_W^r} \tag{7}$$

其中，类间方差 v_B 可表示为式(8)：

$$v_B^r=\frac{1}{k}\sum_{w=1}^{C}k_w(\mu_{wr}-\mu_r)^2 \tag{8}$$

类间方差 v_W 可表示为式(9)：

$$v_W^r=\frac{1}{k}\sum_{w=1}^{C}k_w\sigma_{wr}^2 \tag{9}$$

(2)冗余性度量

由于 χ^2 检验可度量两个随机变量的关联度，因此在进行特征选择时，将其作为特征的冗余性度量。假设特征集中存在两个特征，其随机变量表示为 X，X'，则 χ^2 分布表示为式(10)：

$$\chi^2(X,X')=\sum_{i=1}^{k}\frac{(F_i-F_i')^2}{F_i'} \tag{10}$$

其中，F_i，F_i'分别为两个特征随机变量在第 i 个区间的实际频度和理论频度。假设 X，X'相互独立，若 $p(\chi^2)>\beta$，则称在显著水平 β 下拒绝该假设检验，即 X，X'统计相关，由此可断言两个特征中有一个为冗余特征。保留重要性度量较高的特征完成特征选择，算法如表 5 所列。

表 5　基于 Fisher 比和 χ^2 检验的特征选择算法

输入：原始训练集、候选特征集 Φ、阈值 α、阈值 β	
输出：特征集 Φ′	
Step1	计算特征集 Φ 中各特征的 Fisher 比，并按降序排列，得到候选特征列表 list；
Step2	剔除 list 中 Fisher 比值小于 α 的特征；
Step3	按照由大到小的顺序，依次选取 list 中特征值作为 X，计算该特征值与剩余特征的 $p(\chi^2)$；
Step4	若 $p(\chi^2)>\beta$，则剔除该特征；
Step5	选取下一个特征，重复 Step3、Step4，直至所有特征都经过 χ^2 检验；
Step6	将经过选择的 list 输出为 Φ′。

通过特征选择，能够降低不相关特征的干扰，减少冗余特征的计算，从而进一步提高社会关系判断性能。

2.3　随机森林算法

目前，在机器学习领域，用于分类的算法数不胜数，文献[17]采用 121 个数据集，对 17 大类的 179 类算法进行对比实验。结果表明，平均精度在 Top20 的算法中排名最靠前的为

随机森林、SVM。考虑到随机森林训练速度快，实现简单，不仅能够有效克服过拟合的问题，还能够检测到特征之间的互相影响[18]，本文拟采用该随机森林对以上特征进行学习，以实现对社会关系的判断。随机森林算法训练流程如表 6 所列。

表 6　随机森林算法步骤

输入：原始训练集(样本数为 k，特征数为 n)	
输出：k_{sub} 个 CART 树	
Step1	采用 Bootstrap 策略，有放回地随机抽取 k_{sub} 个自助样本集；
Step2	选取一个自助样本集作为根节点，以完全分裂的方式进行训练；
Step3	从 n 个特征中随机选取 m_{try} 个特征($m_{try} \leqslant n$)，以节点不纯度(Gini 不纯度、熵不纯度)最小为原则，选取最优特征对节点进行分裂生长；
Step4	最大限度地使节点进行分裂生长，若节点具有最小的不纯度，则将其标记为叶子节点；
Step5	重复 Step2—Step4，直到所有节点都被训练过或被标记为叶子节点；
Step6	重复 Step2—Step5，直到所有的 CART 都被训练过。

训练完成后，将未知样本作为输入，根据构造好的每个 CART 树分类器的投票结果决定样本的分类。

3　实验准备与结果分析

3.1　实验准备

为检验基于时空共现理论进行社会关系推断的可行性，同时考察采用随机森林进行分类判断的有效性，采用带标签的 LBSN 用户的 Check-in 数据进行验证。实验选用 Foursquare 的 Check-in 数据[19]，其基本参数如表 7 所列。

表 7　Foursquare 数据参数

数据编号	采集时间	用户数量	Check-in 数量	关注关系
Data2	2011.10—2011.12	11326	1385223	47164

实验主要分为 3 个部分：1)考察不同的时空共现区对判断结果的影响，以确定恰当的 l 及 T 值；2)分析候选特征集 Φ 与特征集 Φ' 对判断准确率的影响，以验证特征选择方法的有效性；3)对比经过特征选择后的随机森林判断方法与以原始特征为训练集的 AdaBoost、SVM 方法的准确度及计算量，进一步证明基于随机森林的社会关系判断方法的有效性。

实验设 l=500m，T=200min，共从中搜索确定了 573 个时空共现区，具体参数如表 8 所列。

表 8　实验数据参数

区域时间间隔大小 T(min)	区域空间网格大小 l(m)	时空共现区数目	两用户在时空共现区的总数目	关注关系	非关注关系
200	500	573	896	108	788

实验根据每一条数据 D_j 计算出 46 种特征，并以此作为机器学习的输入，其中 $D_j=(u_m,u_n,L_i)$，$j=1,2,\cdots,896$。实验中，RF 的 CART 树的数目取 550 个，SVM 采用 libSVM 工具箱(RBF 核)，AdaBoost 采用 GML AdaBoostMatlab Toolbox，其他实验环境如表 9 所列。

表 9　实验环境

	参数	参考值
硬件环境	CPU	Intel(R) Core(TM) i5-3320M
	主频	2.6GHz
	内存	6GB
软件环境	OS	Windows 8.1 Pro
	软件	MATLAB 7.0

3.2　结果分析

3.2.1　时空共现区对判断结果的影响分析

由图 2 可知，随着 l 及 T 的增加，两用户同在时空共现区的次数增加，从而使得数据量增加。为考察 l 及 T 对社会关系判断的影响，l 及 T 分别取不同值，划分不同的时空共现区，并提取数据、计算特征作为随机森林的输入，图 5、图 6 给出 l、T 与社会关系判断准确率的关系。

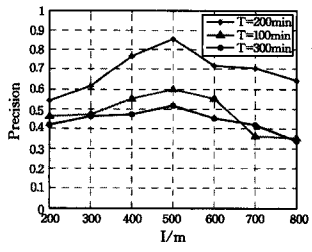


图 5　l 与社会关系判断准确率的关系

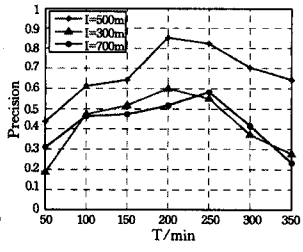
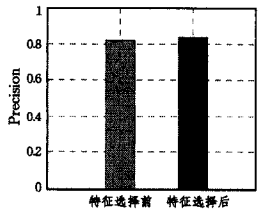


图 6　T 与社会关系判断准确率的关系

由图 5、图 6 可看出，随着 l 及 T 的增加，社会关系判断准确率呈先升高后降低的变化趋势，这是由于当时空共现区设定较小时，用户签到位置具有较大的误差，从而容易导致本应属于同一区域的用户落到不同区域，影响了特征计算。此外，由于提取的数据过少，相应的特征较小，区分度过小，从而使得 CART 树在节点分裂时更加困难。而当 l 及 T 较大时，较大的时空共现区包含了过多的样本点，这使得时空共现区失去了意义，从而影响了社会关系的判断。由此可知，l 及 T 的取值是影响社会关系判断的准确率的一个前提因素，不合理的取值无法获得准确的判断效果。

3.2.2　特征选择对判断结果的影响分析

分别采用未经特征选择的 Φ 及 Φ'，并采用随机森林算法进行分类判断，统计了特征选择前后社会关系判断准确率及单样本训练时间，如图 7、图 8 所示。



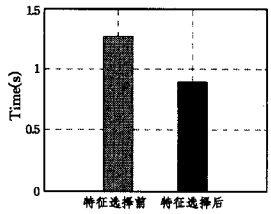


图 7　特征选择前后准确率对比　图 8　特征选择前后运行时间对比

由图 7 可以看出，经过特征选择后，社会关系判断的准确率略高。这是由于在特征选择环节，剔除了冗余和不相关特征，从而减少了这些特征对节点分裂所造成的干扰，增强了随机森林的学习强度，降低了随机森林各树之间的相关性，最终提高了社会关系判断的准确率。由图 8 可以看出，经过特征选择的随机森林判断方法的计算量要远低于未进行特征选择的判断方法，这主要是由于特征空间维度的降低减少了每一个分裂节点不纯度的计算量。总结来看，特征选择能够在保证预测准确率不降低的前提下，有效降低分类判断的计算量。

3.2.3　不同判断方法的结果对比分析

考察了原始特征集为输入、采用 AdaBoost、SVM 的判断方法和所提出的采用特征选择和随机森林的判断方法的性能。如图 9—图 11 所示，重复进行 1000 次实验，统计了 3 种不同方法的准确率、召回率和单样本训练时间。

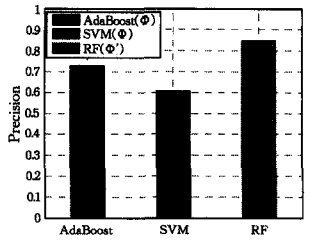


图 9 不同判断方法准确率对比

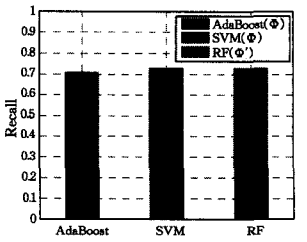


图 10 不同判断方法召回率对比

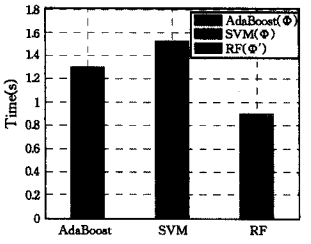


图 11 不同判断方法计算量对比

图 9、图 10 所示为 AdaBoost(Φ)，SVM(Φ)，RF(Φ′) 3 种方法的准确率及召回率对比。由此可看出，与 AdaBoost 及 SVM 相比，随机森林在判断社会关系上具有较高的准确率。这主要是由于其不仅能够较好地克服 over-fitting 的问题，且能够更好地处理多维变量相关的问题。

由图 11 可看出，将 Filter 特征选择与随机森林相结合的方法比采用原始特征和其他算法(AdaBoost、SVM)的社会关系判断方法具有更低的计算量，具体原因已在 3.2.2 节中说明，此处不再赘述。由此可见，所提方法能够以较低的计算代价和较高的准确率实现社会关系推断。

结束语 利用时空共现理论及数据挖掘方法从 LBS 用户的时空信息中推断其是否存在社会关系是基于位置大数据的情报挖掘中的一个新兴分支。在引入时空共现这一概念的基础上，总结归纳了 4 类特征，采用一种 Fisher 准则和 χ^2 检验进行特征选择，将随机森林方法应用到社会关系推断中。分别研究了社会关系判断结果与时空共现区的大小、是否经过特征选择、不同机器学习方法之间的关系。实验结果表明：1)丰富的数据集、合适的 l 及 T 是确保社会关系判断可行的前提；2)基于 Fisher 比和 χ^2 检验的特征选择方法能够有效剔除特征集中的不相关特征和冗余特征；3)相比于其他方法(SVM、AdaBoost)，采用特征选择和随机森林相结合的社会关系判断方法具有更低的计算代价和更高的准确率。今后将以此为基础进一步研究社会关系类型、社会网络的度与本文所选择的特征之间的关联性。

参 考 文 献

[1] Zickuhr K. Location-based services [EB/OL]. (2013-09-12) [2016-11-14]. http://www.pewintemet.org/2013/09/12/location-based.services

[2] Lu X, Wetter E, Bharti N, et al. Approaching the limit of predictability in human mobility[J]. Scientific Reports, 2013, 3(10):1-9

[3] Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility[J]. Science, 2010, 327(5968):1018-1021

[4] Guo C, Liu J N, Fang Y, et al. Value extraction and collaborative mining methods for location big data[J]. Journal of Software, 2014, 25(4):713-730 (in Chinese)
郭迟，刘经南，方媛，等. 位置大数据的价值提取与协同挖掘方法[J]. 软件学报，2014，25(4)：713-730

[5] Psorakis I, Voelkl B, Garroway C J, et al. Inferring social structure from temporal data[J]. Behavioral Ecology and Sociobiology, 2015, 69(5):857-866

[6] Psorakis I, Roberts S J, Rezek I, et al. Inferring social network structure in ecological systems from spatio-temporal data streams[J]. Journal of the Royal Society Interface, 2012, 9(76):3055-3066

[7] Jayadevan V, Bharadwaj K, Kumar A, et al. Discovering Local Social Groups using Mobility Data[J]. International Journal of Computer Applications, 2015, 120(21):15-20

[8] Lim K H, Chan J, Leckie C, et al. Detecting location-centric communities using social-spatial links with temporal constraints[C]// European Conference on Information Retrival. 2015:489-494

[9] Yu R, He X, Liu Y. GLAD: group anomaly detection in social media analysis[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014:1-7

[10] Steurer M, Trattner C. Acquaintance or partner? Predicting partnership in online and location-based social networks[C]// 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). 2013:372-379

[11] Crandall D J, Backstrom L, Cosley D, et al. Inferring social ties from geographic coincidences[J]. Proceedings of the National Academy of Sciences, 2010, 107(52):22436-22441

[12] Li R, Liu J, Yu J X, et al. Co-occurrence prediction in a large location-based social network[J]. Frontiers of Computer Science, 2013, 7(2):185-194

[13] Cranshaw J, Toch E, Hong J, et al. Bridging the gap between physical location and online social networks[C]// Proceedings of the 12th ACM International Conference on Ubiquitous Computing. 2010:119-128

[14] Hsieh H, Yan R, Li C. Where You Go Reveals Who You Know: Analyzing Social Ties from Millions of Footprints[C]// Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. ACM, 2015:1839-1842

[15] Shokri R, Theodorakopoulos G, Danezis G, et al. Quantifying location privacy: the case of sporadic location exposure[C]// Proceedings of the 11th International Symposium PETS 2011. 2011:57-76

[16] Cho E, Myers S A, Leskovec J. Friendship and mobility: user movement in location-based social networks[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011:1082-1090

[17] Fernández-Delgado M, Cernadas E, Barro S, et al. Do we need hundreds of classifiers to solve real world classification problems? [J]. The Journal of Machine Learning Research, 2014, 15(1):3133-3181

[18] Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: A survey and results of new tests[J]. Pattern Recognition, 2011, 44(2):330-349

[19] Gao H, Tang J, Liu H. Exploring Social-Historical Ties on Location-Based Social Networks[C]// The 6th International Aaai Conference on Weblogs And Social Media. 2012:1-8