软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(10):2533−2549 [doi: 10.3724/SP.J.1001.2012.04228]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种上下文移动用户偏好自适应学习方法*

史艳翠[1,2+], 孟祥武[1,2], 张玉洁[1,2], 王立才[1,2]

[1](智能通信软件与多媒体北京市重点实验室(北京邮电大学),北京 100876)

[2](北京邮电大学 计算机学院,北京 100876)

Adaptive Learning Approach of Contextual Mobile User Preferences

SHI Yan-Cui[1,2+], MENG Xiang-Wu[1,2], ZHANG Yu-Jie[1,2], WANG Li-Cai[1,2]

[1](Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia (Beijing University of Posts and Telecommunications), Beijing 100876, China)

[2](School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China)

+ Corresponding author: E-mail: shi_yancui@126.com, http://www.bupt.edu.cn

Shi YC, Meng XW, Zhang YJ, Wang LC. Adaptive learning approach of contextual mobile user preferences. *Journal of Software*, 2012,23(10):2533−2549 (in Chinese). http://www.jos.org.cn/1000-9825/4228.htm

Abstract: A mobile network has higher demands for the performance of personalized mobile network services, but existing researches have been unable to modify the contextual mobile user preferences adaptively and provide real-time, accurate personalized mobile network services for mobile users. This paper proposes a context computing-based approach to mobile user preferences adaptive learning, which can ensure the accuracy and the response time. First, through analyzing the logs of contextual mobile user behaviors, the method judges whether mobile user behaviors are affected by context or not, and detects whether the contextual mobile user behaviors change. According to these judgments, the contextual mobile user preferences are modified. Secondly, the context is introduced into the least squares support vector machine (LSSVM), which is employed to learn the changed contextual mobile user preferences. Further, a learning method of contextual mobile user preferences is proposed which is based on context of the least squares support vector machine (C-LSSVM). Finally, the experimental results show that the proposed method is superior to other learning methods when considering both accuracy and response time. The proposed method in this paper can be applied in the system of personalized mobile network services.

Key words: mobile network; preference learning; contextual mobile user preferences; C-LSSVM

摘 要: 针对移动网络对个性化移动网络服务系统的性能提出了更高的要求,但现有研究难以自适应地修改上下文移动用户偏好以为移动用户提供实时、准确的个性化移动网络服务的问题,提出了一种上下文移动用户偏好自适应学习方法,在保证精确度的基础上缩短了学习的响应时间.首先,通过分析移动用户行为日志来判断移动用户行为是否受上下文影响,并在此基础上判断移动用户行为是否发生变化.然后,根据判断结果对上下文移动用户偏好进行修正.在对发生变化的上下文移动用户偏好进行学习时,将上下文引入到最小二乘支持向量机中,进一步提出了基于上下文最小二乘支持向量机(C-LSSVM)的上下文移动用户偏好学习方法.最后,实验结果表明,当综合考虑精确度和

* 基金项目: 国家自然科学基金(60872051); 中央高校基础研究基金(2009RC0203); 北京市教育委员会共建项目专项资助
收稿时间: 2011-04-06; 修改时间: 2011-08-31, 2012-02-21; 定稿时间: 2012-04-01

响应时间两方面因素时,所提出的方法优于其他学习方法,并且可应用于个性化移动网络服务系统中.

关键词: 移动网络;偏好学习;上下文移动用户偏好;上下文最小二乘支持向量机

中图法分类号: TP181 文献标识码: A

随着下一代网络技术的飞速发展,移动通信网络在与计算机网络逐渐融合的过程中,对互联网信息服务进行了延伸,为移动用户提供了比传统通信业务更加丰富多彩的移动网络服务[1],而且这些移动网络服务在内容、价格、QoS(quality of service,服务质量)等方面也存在较大差异[2].与此同时,由于 3G 网络的商用、智能移动设备的日益普及,信息资源的获取和推送可以发生在“任何时间、任何地点、以任何方式”,为移动用户提供无处不在的移动网络服务已经成为可能[3].然而,随着移动网络服务的日益涌现及其广泛应用,移动网络服务类型和信息内容的增长将逐渐超出人们所能接受的范围,加之移动设备的界面显示、终端处理、输入/输出等能力有限,将导致严重的“移动信息过载”问题[4]的出现.为了准确提供和推荐移动用户真正感兴趣的移动网络服务及其信息内容,个性化移动网络服务研究成为学术界和工业界近年来的研究热点.

与传统计算机网络相比,移动用户面临更加复杂多变的信息服务提供环境,各种类型的上下文信息对移动用户信息需求的影响更加明显.为了及时、准确地为移动用户提供个性化的移动网络服务,上下文移动用户偏好提取技术也成为近年来的研究热点之一[5].移动用户信息需求总是不断变化的,上下文约束下的移动用户偏好往往也随着时间的迁移而动态变化,这就需要上下文移动用户偏好学习方法能够及时捕捉和响应上述变化,自适应获取实时、精确、无冲突的上下文移动用户偏好.目前,大多数研究采用定期重新提取上下文移动用户偏好的办法对其进行更新.然而上下文移动用户偏好可能只有部分发生变化,而且移动用户数量多,重新提取上下文移动用户偏好的计算复杂度较大,不能满足移动用户实时、准确的个性化信息需求.

针对上述问题,本文提出了一种上下文移动用户偏好自适应学习方法.该方法通过分析上下文移动用户行为日志来判断上下文移动用户行为是否发生变化.当上下文移动用户行为不发生变化时,只对相应的上下文移动用户偏好的可信度进行修正;当上下文移动用户行为发生变化时,采用分类方法进行学习.由于只对部分上下文移动用户偏好进行学习,加快了学习的响应时间.为了保证上下文移动用户偏好的准确性并进一步加快其学习的响应时间,本文将上下文引入到最小二乘支持向量机分类方法中,并提出了一种基于上下文最小二乘支持向量机的上下文移动用户偏好学习方法,以对发生变化的上下文移动用户行为进行学习.

本文第 1 节主要介绍上下文用户偏好提取方法和用户偏好自适应模型的相关工作.第 2 节引入上下文最小二乘支持向量机.第 3 节介绍本文提出的上下文移动用户偏好学习方法.第 4 节给出实验结果和相关结论.第 5 节总结本文工作并给出下一步的研究方向.

1 相关工作

近几年,国内外对基于上下文感知的用户偏好获取技术进行了大量研究.用户偏好的表示可分为定性和定量两种表示方法[4]:定性偏好主要考察两个偏好之间的二元偏序关系,从逻辑推理和偏序模型的角度来表示,例如文献[6]中通过构建逻辑链来表示多个偏好之间的关系;定量偏好通过效用函数和数字评分量化表示.常用的上下文移动用户偏好提取技术有贝叶斯网络[7]、神经网络[8]、关联规则[9]、决策树[10]以及基于本体[11]的方法等.上述方法各有其优缺点,基于贝叶斯网络、决策树的提取方法较为简单,但方法能力较弱;基于神经网络的提取方法有局部极小点、过学习以及结构和类型的选择过分依赖于经验等固有的缺陷;基于本体的提取方法在本体设计时很大程度上依赖于研究人员的知识和经验,当公理设置不准确时,得到的上下文移动用户偏好准确率较低;基于关联规则的提取方法计算复杂度较大.

随着时间的迁移,部分上下文移动用户偏好会发生变化.为了给移动用户提供实时、准确的个性化服务,需要对发生变化的上下文移动用户偏好及时进行自适应更新.用户偏好自适应学习方法包括两种:显性自适应学习和隐性自适应学习.显性自适应学习是指根据用户反馈信息来更新用户偏好集[10],如用户评分;隐性自适应学习通过监测和机器学习来更新用户偏好[11].文献[5]考虑到上下文对用户偏好的影响,通过分析上下文关联的用

户行为来挖掘用户偏好,但没有考虑上下文用户偏好会随时间发生变化的问题,个性化服务的实时性受到影响.文献[12]针对个性化信息服务难以有效适应用户需求变化的问题,从行为动机的角度分析了需求变化的原因,通过分析用户行为记录和监测用户反馈,隐性判断其需求变化趋势并主动做出适应,但文献[12]没有考虑上下文因素对用户偏好自适应的影响,因此,个性化服务的精确性也将受到影响.文献[13]根据用户显性反馈信息和隐性反馈信息来更新用户偏好.其中,基于隐性反馈信息的用户偏好更新方法通过分析用户观看电视节目的行为变化来更新用户偏好集.当用户偏好数据量较大时,为避免重新学习时计算复杂度较大的问题,文献[9]中将用户历史行为数据存储在两个数据库中:长期用户行为数据库和临时用户行为数据库.长期用户行为数据库存储所有数据,临时用户行为数据库存储上次用户偏好更新后到现在的数据.当新的更新周期被触发时,在临时用户行为数据集上采用决策树算法重新提取用户偏好.

鉴于上下文移动用户偏好提取技术研究领域的现有工作都存在一些缺点,本文提出一种基于上下文计算的移动用户偏好自适应学习方法.为了及时、准确地对上下文移动用户偏好进行自适应学习,需要判断移动用户偏好受哪些上下文影响以及每种上下文对移动用户偏好的影响程度;随着时间的迁移,上下文移动用户偏好可能会部分地发生变化,当上下文移动用户行为比较稳定时,发生变化的上下文移动用户偏好比较少,因此可以通过检测变化来对其进行修正;当对变化的上下文移动用户偏好进行学习时,为了确保精确度和响应时间,需要选择合适的机器学习方法及时、准确地对其进行学习.

2 上下文最小二乘支持向量机的引入

2.1 上下文移动用户偏好表示

定义 1(上下文移动用户偏好模型). M 是一个五元组,即 $M=\{U,C,S,V,P\}$,其中,U 表示移动用户,C 表示影响移动用户偏好的上下文,S 表示移动用户使用的移动网络服务,V 表示移动用户对移动网络服务的使用量,P 表示移动用户对移动网络服务的偏好.

上下文可以表示为一组特殊属性的集合: $C=\{C_1,C_2,...,C_{n_1}\}$, n_1 表示上下文的种类数.不同的上下文类型在量化时有不同的量化方法,例如:

$$\left.\begin{array}{l} C_1: Time(\text{上午}:1,\text{下午}:2,\text{晚上}:3,\text{深夜}:4) \\ C_2: Location(\text{在家}:1,\text{工作}:2,\text{其他地方}:3) \end{array}\right\} \tag{1}$$

本文中,上下文移动用户偏好采用定量偏好表示方法.定量偏好可以表示为 $P=\{P_{preference},P_{cre}\}$,其中,$P_{preference}\in Z$ 表示移动用户对移动网络服务的偏好值,$P_{cre}\in[0,1]$表示偏好值的可信度.

2.2 上下文最小二乘支持向量机

支持向量机(support vector machine,简称 SVM)与其他分类方法相比具有较高的分类精度,但其计算复杂度较大.为了提高标准支持向量机的训练效率,Suykens 等人[14]修改了 SVM 的损失函数和约束条件,使训练转换为线性问题,由此构建的最小二乘支持向量机(least squares support vector machine,简称 LSSVM)不仅能够保持标准 SVM 的泛化能力,而且还可以直接用于多分类问题.文献[15]分析了标准支持向量机、最小二乘支持向量机、朴素贝叶斯(naïve Bayesian,简称 NB)、C4.5 决策树等分类方法的精确度.其中,最小二乘支持向量机比朴素贝叶斯高出 13%,比 C4.5 决策树高出 10.9%,但比标准支持向量机低 1.7%.由此可知,最小二乘支持向量机分类方法在分类精度方面略低于标准支持向量机,但优于朴素贝叶斯、C4.5 决策树等分类方法.

最小二乘支持向量机分类方法不仅具有分类精度方面的优势,其响应时间也较小.表 1 给出了几种典型分类方法的时间复杂度.其中,N 表示训练样本数;m 表示测试样本数,且 $\frac{N}{m}=l$;n 表示样本属性个数;h 表示用户偏好的类型数.

Table 1 Time complexity of several classification methods

表 1 几种分类方法的时间复杂度

	训练时间复杂度	测试时间复杂度
标准SVM	$O(hN^2)$	$O(hm^2)$
LSSVM	$O(n^2N)$[16]	$O(mn^2)$
C4.5	$O(n\text{N}\log N)$	$O(nm)$
NB	$O(nN)$	$O(hnm)$

由于训练样本数和测试样本数一般大于样本属性个数 n 和偏好类型数 h,即 $N>h,N>n$ 且 $m>h,m>n$,可以得到 $O(hN^2)>O(n^2N),O(hm^2)>O(n^2m)$.由此可知,$O(hN^2+hm^2)>O(n^2N+n^2m)$,LSSVM 的学习时间复杂度小于标准 SVM 的学习时间复杂度.

由于样本属性个数较少,因此,$1<n<<N$,即 $n<\log N$,可得 $O(nN\log N)>O(n^2N)$.即,与 LSSVM 相比,C4.5 的训练时间比较长,由表 1 可知,$O(nN\log N+nm)>O(nm(l\log N+1)),O(n^2(N+m))=O(nm(nl+n))$.又因为 $n<\log N,1<n<<N$,可知 $\log N>n+1$,且一般情况下,$l\geqslant 4$.因此,当 $n\leqslant 5$ 时,LSSVM 的学习时间复杂度会小于 C4.5 的学习时间复杂度.本文中,$l=5,n=4$,所以,LSSVM 的学习时间复杂度小于 C4.5 的学习时间复杂度.

当 $n=1$ 时,LSSVM 的学习时间复杂度为 $O(N+m)$,NB 的学习时间复杂度为 $O(N+hm)$.由此可得,当样本属性为 1 时,LSSVN 的复杂度小于 NB 的时间复杂度.当 $n>1$ 时,由表 1 可知,$O(n^2N)>O(nN)$,又因为 h,n 的值相差不大,所以有 $O(n^2(N+m))=O(nN+hnm)$.因此,LSSVM 的学习时间复杂度大于 NB 的学习时间复杂度.

因此可知,LSSVM 分类学习方法的响应时间小于标准 SVM 分类学习方法和 C4.5 分类学习方法的响应时间,但大于 NB 分类学习方法的响应时间.

综上所述,采用最小二乘支持向量机分类方法对用户偏好进行学习,既保证了精确度又加快了学习的响应时间.为了提高上下文移动用户偏好自适应学习的准确度,将上下文引入到最小二乘支持向量机中.

定义 2(上下文最小二乘支持向量机). 上下文最小二乘支持向量机是最小二乘支持向量机的扩展,通过将上下文添加到最小二乘支持向量机的特征向量空间来实现.其样本特征可以表示为 $x_i=(u_i,c_i,s_i,v_i),i=1,2,\ldots,N$,其中,$u_i\in U;c_i\in C$ 且 $c_i=\{c_{1i},c_{2i},\ldots,c_{ti},\ldots,c_{n_ti}\},c_{ti}\in C_t$;$s_i\in S;v_i\in V;x_i$ 表示移动用户在上下文约束下使用移动网络服务行为的第 i 个实例;$y_i=p_j,p_j\in P_{preference}$,表示移动用户第 i 个实例的偏好值.

与最小二乘支持向量机相比,上下文最小二乘支持向量机的样本属性个数 n 会有所增加,因此其计算复杂度会随 n 值正比例增加.由于其他分类方法(NB,C4.5)的计算度同样与样本属性个数 n 成正比例增加,因此上面的理论分析仍然适用于上下文最小二乘支持向量机.由此可知,与其他分类相比,上下文最小二乘支持向量机更适应移动网络的需求.

3 上下文移动用户偏好学习方法

移动网络中,移动用户以及移动网络服务种类比较多,因此,上下文移动用户偏好更新时的计算复杂度比较大.为了及时、准确地为移动用户提供个性化的移动网络服务,对上下文移动用户偏好自适应学习的结果在精确度和响应时间两方面提出了更为严格的要求.随着时间的迁移,上下文移动用户偏好只有部分发生变化.例如,使用移动套餐的用户,由于免费的接听时间、短信条数、上网流量等是有额度的,移动用户一般会将自己的使用量控制在额度范围内,这些偏好一般不发生变化.因此,可以对移动用户行为日志进行分析,先判断其上下文移动用户行为是否发生变化,然后再根据判定结果对上下文移动用户偏好进行修正.在采用本文方法进行学习时,其计算复杂度为 $O(n^2N+m+n^2m_1),m_1$ 为发生变化的上下文移动用户偏好数目.当上下文移动用户偏好比较稳定时,即 $m_1=0$,只需对上下文移动用户偏好的可信度进行修正,计算复杂度为 $O(m)$.当上下文移动用户偏好变化比较频繁时,即 $m_1=m$,需要对所有上下文移动用户偏好进行学习,其计算复杂度为 $O(n^2N+m+n^2m)$.当 $\frac{m-m_1}{m}>\frac{1}{n^2}$ 时,本文方法的计算复杂度小于直接学习的计算复杂度.本文中 $n=4$,因此,当不发生变化的上下文移动用户偏好的个数大于总的上下文移动用户偏好个数的 6.25%时,采用本文方法的计算复杂度比直接学习的复

杂度要小.

3.1 上下文移动用户行为变化检测方法

随着时间的迁移,上下文移动用户偏好会发生变化,上下文对移动用户偏好的影响也会发生变化.因此,上下文移动用户偏好变化检测分为两步:首先要确定上下文对移动用户偏好的影响是否发生变化,然后再检测上下文约束下的移动用户偏好是否发生变化.上下文移动用户偏好的变化可以通过分析相应的上下文移动用户行为的变化获得.文献[13]针对用户对电视内容的偏好,在更新用户偏好的过程中,通过分析用户行为,根据其使用服务时长的变化来判断其相应的偏好是否发生变化.移动网络中,移动用户使用移动网络服务的度量方式不同,例如,语音通信按时长计算、短信按次数计算、数据业务按流量计算等.因此,在计算上下文移动用户偏好时,需要将各种度量方法进行归一化.另外,有些服务不仅需要考虑使用时长,还需要考虑使用次数.例如,语音通信服务,假设移动用户 *A* 在一个月内只使用 1 次,使用时长为 100 分钟;移动用户 *B* 在一个月内使用 10 次,每次 10 分钟,虽然移动用户 *A* 和移动用户 *B* 使用语音服务的时长一样,但移动用户 *B* 对语音服务的偏好要高于移动用户 *A* 对语音服务的偏好.为了更精确地检测上下文移动用户行为是否发生变化,本文在计算移动用户对移动网络服务使用量的变化时,不仅考虑了移动用户使用移动网络服务的时长,还考虑了移动用户使用移动网络服务的次数.

3.1.1 确定影响移动用户行为的上下文

确定移动用户行为受哪些上下文影响,可以根据移动用户在各类单维上下文约束下对移动网络服务使用量的波动率来判断.

定义 3(上下文移动用户行为波动率). 上下文移动用户行为波动率是指上下文约束下,移动用户对某种移动网络服务的使用量的变化程度.上下文移动用户行为波动率可以表示为[5]

$$vol_{t,s_j}=\frac{\frac{1}{n_t}\sum_{c_{tk}\in C_t}|Volume(u_i,s_j,c_{tk})-\overline{Volume(u_i,s_j,C_t)}|}{\overline{Volume(u_i,s_j,C_t)}} \tag{2}$$

其中,

- $s_j\in S$;
- C_t 表示某种类型的单维上下文;
- c_{tk} 表示单维上下文的某个实例;
- n_t 表示上下文 C_t 包含的实例个数;
- $Volume(u_i,s_j,c_{tk})\in V$ 表示移动用户在指定上下文实例约束下对移动网络服务 s_j 的使用量;
- $\overline{Volume(u_i,s_j,C_t)}=\frac{1}{n_t}\sum_{c_{tk}\in C_t}Volume(u_i,s_j,c_{tk})$ 表示移动用户在指定上下文约束下对移动网络服务的平均使用量.

波动率越大,说明移动用户行为受该上下文的影响越大;若所有单维上下文的波动率全部小于设定的阈值,则说明移动用户行为不受上下文影响;如果部分单维上下文的波动率大于设定的阈值,则说明移动用户行为受部分上下文影响.例如在时间上下文约束下,某移动用户对语音通信服务的使用量为(40,60,80,3),说明该移动用户在深夜使用语音通信服务较少,在晚上使用语音通信服务较多,其波动率为 0.53,说明该移动用户使用语音通信服务时受时间上下文影响较大;在位置上下文约束下移动用户对语音通信服务的使用量为(30,27,32),其波动率为 0.06,说明该移动用户使用语音通信服务时受位置上下文影响较小.

当 $vol_{t,s_j}<Vol_{threshold}$ 时,移动用户行为不受上下文 C_t 影响.其中,$Vol_{threshold}$ 为判断移动用户行为是否受上下文影响的波动率阈值,需要根据实际情况加以设定.另外,上下文移动用户偏好的数目与上下文的种类数之间是正比例关系.因此,当 $Vol_{threshold}$=0 时,对波动率的限制比较严格,影响移动用户行为的上下文比较多,学习得到的上下文移动用户偏好的精确度比较高,但学习得到的上下文移动用户偏好比较多,响应时间比较长;当 $Vol_{threshold}=vol_{t,s_j}$ 时,移动用户行为不受上下文影响,学习得到的上下文移动用户偏好比较少,响应时间比较小,

但学习得到的上下文移动用户偏好的精确度比较差.因此,$Vol_{threshold}$ 可以根据对上下文移动用户偏好精确度和响应时间的实际要求进行设定.

定义 4(边际效用递减理论). 用户对服务的使用量越大,则偏好值越大;但是,随着使用量越来越大,用户偏好值增长的趋势会变慢.本文选用对数函数建立移动用户对移动网络服务的使用量和相应的偏好值之间的关系,可以表示为

$$p_i \propto [\log_{\tau_1} v_i] \tag{3}$$

其中,τ_1 表示底数,其值需要根据数据集的实际分布情况设定,相应的设定原则为:

① 当移动用户对移动网络服务的使用量波动较小时,τ_1 的取值要小些;当移动用户对移动网络服务的使用量波动较大时,τ_1 的取值要大些.例如,上下文移动用户偏好值在[1,5]区间上取值时,数据集 *A* 中最大值为 1 000,最小值为 1,变化范围为 0~999,则 τ_1=5.5;数据集 *B* 中最大值为 1 000,最小值为 800,变化范围为 0~200,则 τ_1=3.5;

② 根据移动用户需求行为的幂律分布特性,80%左右的上下文移动用户偏好值应该处在中间取值的部分.例如,当上下文移动用户偏好值在[1,10]取值时,则 80%左右的上下文移动用户偏好的取值应该在[3,8]区间上,仍以数据集 *B* 为例,则 τ_1=1.75.

3.1.2 上下文移动用户行为变化检测

1) 单维上下文约束下移动用户行为变化检测

根据公式(2)计算移动用户行为在每种上下文约束下的波动率,通过与设定的阈值比较得到影响移动用户行为的上下文,检测在这些上下文约束下的移动用户行为是否发生变化.移动用户对移动网络服务的使用量包括使用次数和使用时长,在计算偏好时,本文考虑了这两种因素,可以表示为

$$Volume(u_i,s_j,c_{tk}) = \upsilon_1 \times \frac{Length(u_i,s_j,c_{tk})}{Length(u_i,c_{tk})} + \upsilon_2 \times \frac{Num(u_i,s_j,c_{tk})}{Num(u_i,c_{tk})} \tag{4}$$

其中,$Length(u_i,s_j,c_{tk})$表示单维上下文约束下移动用户使用移动网络服务 s_j 的时长,$Length(u_i,c_{tk})$表示单维上下文约束下移动用户使用所有移动网络服务总的时长;$Num(u_i,s_j,c_{tk})$表示单维上下文约束下移动用户使用移动网络服务 s_j 的次数,$Num(u_i,c_{tk})$表示单维上下文约束下移动用户使用所有移动网络服务的次数;υ_1 和 υ_2 表示时长和次数的权重值,且 $\upsilon_1+\upsilon_2$=1.由于对时长和次数权重值设定的研究比较少,所以本文通过多次实验选定合适的取值.移动用户对移动网络服务的使用量变化可以表示为

$$Volume(u_i,s_j,c_{tk})_{change} = Volume(u_i,s_j,c_{tk})_{new} - Volume(u_i,s_j,c_{tk})_{old} \tag{5}$$

当 $Volume(u_i,s_j,c_{tk})_{change} > Volume_{threshold}$ 时,表示单维上下文约束下移动用户使用移动网络服务 s_j 的行为发生变化;否则,表示其不发生变化.其中,$Volume_{threshold}$ 表示设定的单维上下文约束下移动用户行为变化的阈值.

2) 多维上下文约束下移动用户行为变化检测

多维上下文约束下移动用户对移动网络服务的使用量可以根据单维上下文约束下移动用户对移动网络服务的使用量来计算,本文采用文献[17]中的方法计算多维上下文约束下移动用户对移动网络服务的使用量.计算公式可以表示为

$$Volume(u_i,s_j,c_k) = \sum_{t=1}^{n_1} \omega_{t,s_j} \times Volume(u_i,s_j,c_{tk}) \tag{6}$$

$$\omega_{t,s_j} = \frac{vol_{t,s_j}}{\sum_{t=1}^{n_1} vol_{t,s_j}} \tag{7}$$

其中,$Volume(u_i,s_j,c_k) \in V$ 表示移动用户在多维上下文约束下对移动网络服务 s_j 的使用量,ω_{t,s_j} 表示单维上下文对移动用户行为的权重值.当 $\sum_{t=1}^{n_1} vol_{t,s_j} = 0$ 时,表示移动用户使用移动网络服务 s_j 的行为不受任何上下文影响,移动用户对移动网络服务的使用量为给定更新周期内总的使用量.多维上下文移动用户行为变化的计算公式可以

表示为

$$Volume(u_i,s_j,c_k)_{change}=Volume(u_i,s_j,c_k)_{new}-Volume(u_i,s_j,c_k)_{old} \tag{8}$$

当 $Volume(u_i,s_j,c_k)_{change}>Volume_{threshold}$ 时,表示多维上下文约束下,移动用户使用移动网络服务 s_j 的行为发生变化;否则,表示该移动用户行为不发生变化,$Volume_{threshold}$ 表示设定的上下文移动用户行为变化的阈值.

$Volume_{threshold}$ 值的设定与 $Vol_{threshold}$ 值的设定类似,需要根据学习得到的上下文移动用户偏好的精确度和响应时间的实际需要加以设定.

根据上述上下文移动用户行为的判断结果可以确定哪些上下文移动用户偏好发生变化.下面需要根据判断结果对相应的偏好值和可信度进行修正.

3.2 上下文移动用户偏好修正方法

根据上下文移动用户行为变化检测的结果,可以对上下文移动用户偏好进行修正.当上下文移动用户行为不发生变化时,只需对相应的可信度进行修改;当上下文移动用户行为发生变化时,采用上下文最小二乘支持向量机分类方法对发生变化的上下文移动用户偏好进行学习.当采用上下文最小二乘支持向量机分类方法时,为了确保学习结果的准确率,需要设定合适的参数,本文通过理论分析确定参数的选取范围,然后通过遗传算法在给定的参数范围内选择最优的参数值.

3.2.1 上下文最小二乘支持向量机分类方法参数设定

上下文最小二乘支持向量机回归多类分类方法如下[18]:对于一个给定的训练数据集$(x_i,y_i),i=1,2,\ldots,N$,构造一个回归函数 $f(x)$,使该函数能够反映 x_i 和它所对应的 y_i 之间的关系.因此,对每一个样本 x_i,使它所对应的回归函数的函数值 $f(x_i)$尽可能地接近 y_i,这样就把多类分类问题转化为函数回归问题.

上下文最小二乘支持向量机回归的优化问题可以表示为

$$\left.\begin{aligned}&\min J(\omega,\xi)=\frac{1}{2}\omega^T\bullet\omega+\gamma\sum_{i=1}^{N}\xi_i^2\\&\text{s.t. } y_i=\omega^T\bullet\psi(x_i)+b+\xi,i=1,2,...,N\end{aligned}\right\} \tag{9}$$

用拉格朗日法求解上述优化问题,上下文最小二乘支持向量机优化问题转化为求解线性方程,可以表示为

$$\begin{bmatrix}0&1&1&1\\1&k(x_1,x_1)+1/\gamma&\cdots&k(x_1,x_N)\\1&\vdots&\ddots&\vdots\\1&k(x_N,x_1)&\cdots&k(x_N,x_N)+1/\gamma\end{bmatrix}\times\begin{bmatrix}b\\a_1\\\vdots\\a_N\end{bmatrix}=\begin{bmatrix}0\\y_1\\\vdots\\y_N\end{bmatrix} \tag{10}$$

其中,$k(x_i,x_j)$表示核函数,γ表示正则参数.设$\Omega_{ij}=k(x_i,x_j)$,则公式(10)可以表示为

$$\begin{bmatrix}0&\vec{1}\\\vec{1}&\Omega+\gamma^{-1}I\end{bmatrix}\times\begin{bmatrix}b\\a\end{bmatrix}=\begin{bmatrix}0\\y\end{bmatrix} \tag{11}$$

令 $A=\Omega+\gamma^{-1}I$,求解公式(11)可以得到:

$$a=A^{-1}(y-b\vec{1}) \tag{12}$$

$$b=\frac{\vec{1}^TA^{-1}y}{\vec{1}^TA^{-1}\vec{1}} \tag{13}$$

因此,可以得到上下文最小二乘支持向量机的回归函数,表示如下:

$$f(x)=\sum_{i=1}^{N}a_ik(x,x_i)+b \tag{14}$$

上下文最小二乘支持向量机多类分类函数可以表示为

$$j^*=\arg\min_{1\leqslant j\leqslant h}\{|f(x)-p_j|\} \tag{15}$$

本文中核函数取径向基核函数,$k(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$,$k(x_i,x_j)$的值可以看作是求解 x_i 和 x_j 相似度的一种方法.在公式(14)中,$k(x,x_i)$为 x 和 x_i 的相似度,a_i 表示相似度的权重值.当对测试数据进行分类时,根据当前数据与

所有训练数据的相似度以及权重值确定其属于的类别.其中,σ表示核函数的宽度.因此,公式(14)中需要确定γ和σ的值.

定理 1. 当 x_i∈训练集,且$\sigma\to 0,\gamma\to\infty$时,$f(x_i)\to y_i$.

证明:当 $\sigma\to 0$ 时,$k(x_i,x_i)\to 1$,若 $j\neq i,k(x_i,x_j)\to 0$.即,当 $\sigma\to 0$ 时,$f(x_i)\to a_i+b$.由公式(10)可知,当 $\sigma\to 0,\gamma\to\infty$时,$a_i+b\to y_i$.可知,当 x_i∈训练集,且$\sigma\to 0,\gamma\to\infty$时,$f(x_i)\to y_i$. □

定理 2. 当 x_i∈测试集,且 $x_i\notin$训练集,$\sigma\to 0$ 时,$f(x_i)\to b$.

证明:当$\sigma\to 0$ 时,$k(x_i,x_j)\to 0$,由公式(14)可知,$f(x_i)\to b$.得证. □

定理 3. 当$\sigma\to\infty$时,$f(x_i)\to b$.

证明:当 $\sigma\to\infty$时,$k(x_i,x_j)\to 1$, $f(x_i)\to\sum_{i=1}^{l}a_i+b$,由公式(10)可知,$\sum_{i=1}^{l}a_i=0$.可以得到:当$\sigma\to\infty$时,$f(x_i)\to b$. □

设定σ值时不仅要考虑训练精度,还要兼顾测试精度.根据定理 1~定理 3,本文中σ的取值要能区分训练集和测试集中各个样本.

推理 1. σ的最佳取值范围为$(0.2\times\max(\|x_i-x_j\|),\max(\|x_i-x_j\|))$.

证明:当 $\frac{\|x_i-x_j\|^2}{2\sigma^2}>13$ 时,$k(x_i,x_j)<0.000001$.即,当$\sigma<0.2\times\|x_i-x_j\|$时,不能区分测试集中的数据.因此,$\sigma$的最小值为 $0.2\times\max(\|x_i-x_j\|)$.

为了兼顾训练精度和测试精度,σ的取值是在能区分测试集的情况下尽量地小.因此,σ的最大值取为

$$\max(\|x_i-x_j\|)$$

综上所述,σ的最佳取值范围为$(0.2\times\max(\|x_i-x_j\|),\max(\|x_i-x_j\|))$. □

定理 4. 当$\gamma\to 0$ 时,$f(x_i)\to b$.

证明:当$\gamma\to 0$ 时,$A^{-1}\to 0$.由公式(12)可知,$a_i\to 0,i=1,2,\ldots,N$.由公式(14)可得:当$\gamma\to 0$ 时,$f(x_i)\to b$. □

定理 5. 当$\gamma\to\infty$且 x_i∈训练集时,$f(x_i)\to y_i$.

证明:当$\gamma\to\infty$时, $A^{-1}\to\Omega^{-1}=\begin{bmatrix}k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_N)\\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_N)\\ \vdots & \vdots & \vdots & \vdots\\ k(x_N,x_1) & k(x_N,x_2) & \cdots & k(x_N,x_N)\end{bmatrix}^{-1}$,

$[k(x_1,x_i)\ k(x_2,x_i)\ \ldots\ k(x_N,x_i)]$与 A^{-1} 中$[k(x_1,x_i)\ k(x_2,x_i)\ \ldots\ k(x_N,x_i)]$对应的向量正交.所以可得:

$$f(x_i)\to[0,\ldots,1,\ldots,0]\begin{bmatrix}y_1-b\\ y_2-b\\ \vdots\\ y_N-b\end{bmatrix}+b\to y_i-b+b\to y_i.$$ □

定理 6. 当 x_i∈测试集,且 $x_i\notin$训练集,$\gamma\to\infty$时,不能得到 $f(x_i)\to y_i$.

当 x_i∈测试集且 $x_i\notin$训练集时,$[k(x_1,x_i)\ k(x_2,x_i)\ \ldots\ k(x_N,x_i)]$与 A^{-1} 中向量不存在正交关系,不能推出:当$\gamma\to\infty$时,$f(x_i)\to y_i$. □

由定理 4~定理 6 可知:γ值越大,训练集分类精度越高,但不是越大越好;当γ增加到一定程度时,其测试精度会降低.本文中,γ取值范围为(100,10000).

在确定了σ和γ的取值范围之后,需要在区间上选择合适的值,使学习得到的上下文移动用户偏好的精确度最好.本文采用常用的基于遗传算法的参数优化方法确定σ和γ在区间上的最优值,其中,遗传算法的适应度函数可以表示为

$$\xi_{MRE}=\frac{\sum_{i=1}^{m}\frac{|f(x_i)-y_i|}{y_i}}{m}\tag{16}$$

其中,m 表示测试集中的样本数.

3.2.2 上下文移动用户偏好修正规则

上下文移动用户偏好修正时分为两种情况:上下文移动用户行为不发生变化时上下文移动用户偏好的修正和上下文移动用户行为发生变化时上下文移动用户偏好的修正.上下文移动用户偏好修正分为两步:① 上下文移动用户偏好值修正;② 上下文移动用户偏好值的可信度修正.上下文移动用户偏好值与上下文约束下移动用户对移动网络服务的使用量有关,上下文移动用户偏好值的可信度与上下文移动用户偏好的稳定度有关.即,上下文移动用户偏好值重复次数越多,其可信度越大.

1) 上下文移动用户行为不发生变化

当上下文移动用户行为不发生变化时,相应的上下文移动用户偏好值也不发生变化,但其对应的可信度需要增加.上下文移动用户偏好的可信度与其重复的次数有关,重复次数越多,其可信度越大.可信度的增长趋势和偏好值的增长趋势类似,符合边际效用递减理论.可信度计算公式可以表示为

$$Cre=\begin{cases}Cre_{old}+\alpha, & N_{repeat}<\theta\\ Cre_{old}+\log_{\tau_2}^{N_{repeat}}-\log_{\tau_2}^{N_{repeat}-1}, & N_{repeat}\geqslant\theta\end{cases} \tag{17}$$

其中,

- $Cre\in P_{cre}$,初始值为 0;
- Cre_{old}表示原来上下文移动用户偏好值的可信度;
- N_{repeat}表示上下文移动用户偏好值的重复次数,当上下文移动用户行为不发生变化时加 1;
- α为常数,表示当偏好重复时,可信度增加的幅度;
- θ表示边界值,当重复次数达到一定程度时,减小可信度增加的趋势;
- τ_2 表示对数底数,其值需要根据实际情况而定:当可信度增加幅度较大时,τ_2 值可以小些;当可信度增加幅度较小时,τ_2 值可以大些.

τ_2,α和θ的取值与上下文移动用户偏好学习的周期有关:当周期较短时,α应该小一些,τ_2 和θ大一些;当周期较长时,α应该大一些,τ_2 和θ小一些.当 Cre_{old}=1 时,可信度不再增加.

2) 上下文移动用户行为发生变化

(1) 上下文移动用户偏好学习

当上下文移动用户行为发生变化时,通过上下文最小二乘支持向量机分类方法对发生变化的上下文移动用户行为进行重新学习,得到新的上下文移动用户偏好.首先,对上下文移动用户历史行为数据和已有上下文移动用户偏好进行训练;然后,根据当前的上下文移动用户行为,学习得到新的上下文移动用户偏好.

(2) 上下文移动用户偏好值可信度修正

- 原有上下文移动用户偏好值可信度修正

由于上下文移动用户偏好值发生变化,原有上下文移动用户偏好值的可信度下降,并且其重复次数减 1,其计算公式可以表示为

$$Cre=\begin{cases}Cre_{old}-\alpha, & N_{repeat}<\theta\\ Cre_{old}-(\log_{\tau_2}^{N_{repeat}}-\log_{\tau_2}^{N_{repeat}-1}), & N_{repeat}\geqslant\theta\end{cases} \tag{18}$$

其中,参数取值与公式(17)中的参数取值需保持一致.

- 学习得到的上下文移动用户偏好值可信度修正

① 遍历原有上下文移动用户偏好集,判断学习得到的上下文移动用户偏好是否为新偏好,如果为新偏好,则对其偏好值的可信度进行初始化,赋予默认值 $Cre_{default}$=0.1;否则,执行步骤②;

② 比较学习得到的上下文移动用户偏好值与其他所有该偏好的偏好值,如果为已有上下文移动用户偏好值,则重复次数加 1,根据公式(17)计算其可信度;否则,执行步骤③;

③ 由于该偏好值为移动用户已有偏好新的偏好值,其可信度与已有偏好其他偏好值的可信度有关.其计算公式可以表示为

$$Cre = Cre_{default} + Cre_0 + \sum_{i=1,i\neq c}^{r} 0.01\times \beta_i \times Cre_i \times \left(1-\frac{abs(p_0-p_i)}{p_i}\right) \tag{19}$$

其中,$abs(\cdot)$为取绝对值函数;r 表示上下文移动用户偏好原有偏好值的个数;Cre_i 表示偏好值的可信度;β_i 表示原有偏好值可信度对现有偏好值可信度的影响趋势,其计算公式可以表示为

$$\left.\begin{aligned} Cre_0 &= \left(1-\frac{abs(p_0-p_c)}{p_c}\right)\times Cre_c \\ abs(p_0-p_c) &= \min_{i\in r} abs(p_0-p_i) \end{aligned}\right\} \tag{20}$$

$$\beta_i = \begin{cases} 1, & abs(p_0-p_i) \leqslant 0.5\times p_i \\ -1, & abs(p_0-p_i) > 0.5\times p_i \end{cases} \tag{21}$$

其中,p_i 表示已有上下文移动用户偏好值,p_0 表示学习得到的上下文移动用户偏好值.

4 实验与结果分析

本节描述实验方案设计以及实验结果分析,实验的硬件环境和系统软件见表 2.

Table 2 Hardware and software of system

表 2 硬件环境和系统软件

操作系统	内存	CPU	开发语言和工具		数据库
Windows XP	2GB	2.8GHz	JDK1.5.0_04, Eclipse3.2	Matlab7.0	MySQL5.0

4.1 实验数据

本文实验首先采用麻省理工学院多媒体实验室 MIT 收集的数据集[19]进行实验,该数据集包括 94 个移动用户从 2004 年 9 月份到 2005 年 6 月份共 9 个月移动用户使用手机的行为和相应的上下文信息.但数据集中不包括上下文移动用户偏好信息,因此,首先对 9 个月的上下文移动用户行为进行学习,提取每个月的上下文移动用户偏好.

至今公开可用的包含上下文信息的移动用户行为数据集比较少,MIT 数据集的缺点是移动用户、移动网络服务、上下文等数量比较少,不利于分析更多数据量的情况.因此,本文在调研分析移动用户行为研究报告和抓取中国移动网上应用商城提供的移动网络服务信息的基础上,设计合理的上下文规则、移动用户历史行为上下文规则以及移动用户行为变化规则进行约束,构造了一个包含上下文信息的模拟数据集 MobileServices.模拟数据包括 500 个移动用户在连续 6 个月内的移动用户行为.模拟数据集中,规则是根据经验设定的,可能与移动用户的实际行为稍有偏差,因此实验以 MIT 数据集为主,模拟数据集作为补充.

本文两个数据集中偏好值的取值范围为[1,10],服务相对使用量的变化范围在[0,2],在运用公式(3)时,将变化范围映射到[0,200]区间上,公式(3)中的 τ_1=1.75.

4.2 基准对比方法

本文第 2 节从理论上分析了上下文最小二乘支持向量机在精确度和响应时间两方面的优势,下面通过实验表明上下文最小二乘支持向量机的有效性.在进行对比实验时,本文选择了一些上述分类算法的改进方法作为基准方法,如下所示:

(1) 标准 SVM.核函数取径向基核函数,正则化参数 γ=2500、不敏感参数 ε=0.001、核函数参数 σ=0.4;

(2) 基于加权的 NB.通过对属性进行加权可以提高分类的精确度;

(3) 基于修剪算法的 C4.5.通过修剪算法可以提高 C4.5 的精确度并能减少计算复杂度.

4.3 评价指标

评价指标采用统计度量方法中的平均绝对误差(mean absolute error,简称 MAE)、正确率(precision)和学习的响应时间[20].

平均绝对误差(MAE):通过计算学习得到的上下文移动用户偏好值与实际的上下文移动用户偏好值之间的偏差来度量上下文移动用户偏好学习的精确度.MAE 越小,上下文移动用户偏好学习的精确度越高.若学习得到的上下文移动用户偏好值表示为 $P'=\{p'_1,p'_2,...,p'_m\}$,实际的上下文移动用户偏好值表示为 $realP=\{q_1,q_2,\ldots,q_m\}$,则平均绝对误差计算公式可以表示为

$$MAE=\frac{\sum_{i=1}^{m}|p'_i-q_i|}{m} \tag{22}$$

正确率(precision):通过学习得到的正确的上下文移动用户偏好个数与总的上下文移动用户偏好个数的比值来度量上下文移动用户偏好学习的精确度.正确率越高,说明学习得到的上下文移动用户偏好的精确度越高.其计算公式可以表示为

$$Precision=\frac{m'}{m} \tag{23}$$

其中,m'表示学习得到的正确的上下文移动用户偏好个数.

4.4 实验步骤与实验结果分析

4.4.1 实验步骤

1) 选择训练集和测试集

为了验证本文方法的有效性,在用 MIT 数据集做实验时,选用前 5 个月的上下文移动用户行为数据作为训练集,第 6 个月的上下文移动用户行为数据作为测试集.由于数据集中某些上下文移动用户行为数据少于 6 个月,因此选出 39 个移动用户的上下文行为数据进行实验.在选用模拟数据集时,选用前 5 个月的上下文移动用户行为数据作为训练集,第 6 个月的上下文移动用户行为作为测试集.

2) 确定影响移动用户行为的上下文

MIT 数据集中包括时间、位置两种上下文;模拟数据集中包括时间、位置、使用设备、活动状况、周围人员这 5 种上下文.根据公式(4)计算移动用户使用移动网络服务的使用量,其中,υ_1,υ_2 设定多组数据对,如表 3 所示.通过实验选取学习结果最好的一组:υ_1=0.3,υ_2=0.7;然后,根据公式(2)计算单维上下文约束下移动用户使用某种移动网络服务的波动率,确定移动用户行为是否受该上下文影响.$Vol_{threshold}$ 的取值为 0,0.02,0.05,0.1,0.15.

Table 3 Different values of υ_1 and υ_2

表 3 υ_1 和 υ_2 的取值

υ_1	0.3	0.4	0.5	0.6	0.7
υ_2	0.7	0.6	0.5	0.4	0.3

3) 单维上下文约束下移动用户行为变化检测

首先,通过步骤 2)得到影响移动用户行为的上下文;然后,利用公式(4)计算在上下文 C_t 约束下移动用户对移动网络服务的使用量;最后,根据公式(5)计算得到在上下文 C_t 约束下移动用户行为的变化量.其中,$Volume_{threshold}=\varsigma_t\times Volume(u_i,s_j,c_{tk})_{old}$,$\varsigma_t$ 取值为 0,0.1,0.2,0.3,0.4.

4) 多维上下文约束下移动用户行为变化检测

首先,利用步骤 2)得到的波动率和公式(7)计算各个上下文的权重值;然后,根据步骤 3)得到单维上下文约束下移动用户行为的变化量和公式(6)计算多维上下文约束下移动用户对移动网络服务的使用量;最后,通过公式(8)计算移动用户使用移动网络服务的变化量.其中,$Volume_{threshold}=\varsigma\times Volume(u_i,s_j,c_k)_{old}$,$\varsigma$ 取值为 0,0.1,0.2,0.3,0.4.

5) 上下文移动用户行为不发生变化时上下文移动用户偏好修正

当上下文移动用户行为不发生变化时,对应的上下文移动用户偏好值不变,只需根据公式(17)对其可信度进行修正,本文中,上下文移动用户偏好学习周期为 1 个月,相关参数设定为 τ_2=10,α=0.1,θ=7.根据上述参数设置,当移动用户对某个移动网络服务的偏好值至少 18 个月保持不变时,可信度接近 1;当移动用户对移动网络服务的偏好 12 个月不变时,其可信度达到 0.9.

6) 上下文移动用户行为发生变化时上下文移动用户偏好修正

(1) 确定上下文最小二乘支持向量机的参数.由于 LSSVM 分类精度与其设定的参数有关,为了得到准确的分类,需要设置合适的参数.根据第 3.2.1 节的理论分析可得:σ的取值范围为(0.01,1.9),γ的取值范围为(100,10000),通过遗传算法在给定的范围区间上选取最优的参数值;

(2) 上下文移动用户行为发生变化时上下文移动用户偏好修正.采用不同的分类方法学习上下文移动用户偏好,并根据公式(18)~公式(21)对相应的上下文移动用户偏好值的可信度进行修正.

4.4.2 实验结果及分析

1) 上下文波动率阈值影响

图 1 为 MIT 数据集中,某移动用户在时间上下文约束下其用户行为受上下文影响的情况.

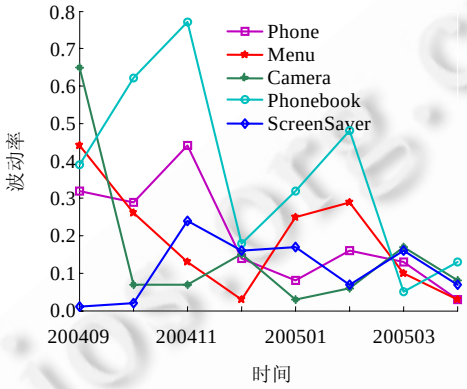


Fig.1 Volatility of a mobile user behaviors under time context (MIT dataset)

图 1 时间上下文约束下某移动用户行为的波动率(MIT 数据集)

从图 1 可以看出,移动网络服务受上下文影响的程度有所不同.Phonebook,Camera 波动率较大,而且变化幅度也较大;ScreenSaver 的波动率较小,而且其波动率变化幅度也较小.这是因为 Phonebook,Camera 这些移动用户的行为受时间影响较大.例如 Camera,由于拍照受亮度影响较大,白天使用照相功能较多,夜晚则较少;ScreenSaver 的波动率较小,是由于当手机不使用时都是处于屏保状态,受时间上下文影响较小.另外,随着时间的迁移,上下文对移动用户行为的影响程度也发生变化.例如 Camera,在 2004 年 9 月的波动率为 0.65,其余月份的波动率都低于 0.2.原因之一是由于移动用户刚使用手机时,感觉拍照比较新鲜,使用比较多,后来失去兴趣,使用量减小,波动率也随之减小.因此,在对上下文移动用户偏好进行更新之前,首先需要确定移动用户行为受哪些上下文影响以及受上下文影响的程度,才能确保上下文移动用户偏好自适应学习的精确度.

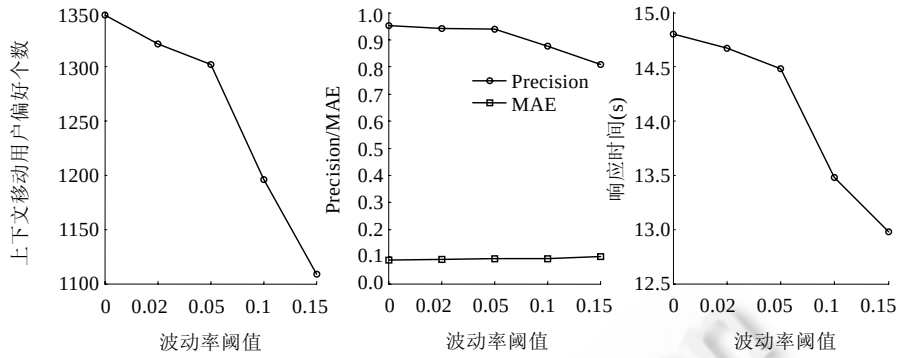
图 2 为在 MIT 数据集中,波动率阈值取不同值时,在时间上下文约束下的移动用户行为通过上下文最小二乘支持向量机分类方法直接学习得到的上下文移动用户偏好精确度比较,其中,LSSVM 的参数为γ=1500,σ=0.17.

由图 2 可知:

① 当 $Vol_{threshold}$ 值增加时,上下文移动用户偏好个数随之减少,当 $Vol_{threshold}$=0.15 时,上下文移动用户偏好个数比 $Vol_{threshold}$=0 时少了 239 个;

② 随着 $Vol_{threshold}$ 值的增加,学习得到的上下文移动用户偏好的精确度随之降低,Precision 降低了大约 15%,而 MAE 增加了 0.012 3.这是由于波动率阈值的限制,使得学习得到的上下文移动用户偏好较少,导致一些受上下文影响的移动用户行为判定为不受上下文影响;

③ 随着 $Vol_{threshold}$ 值的增加,学习响应时间减少.这是因为学习得到的上下文移动用户偏好较少,因此学习响应时间有所降低.综合精确度和响应时间的需求,本文选取的波动率阈值为 0.05.



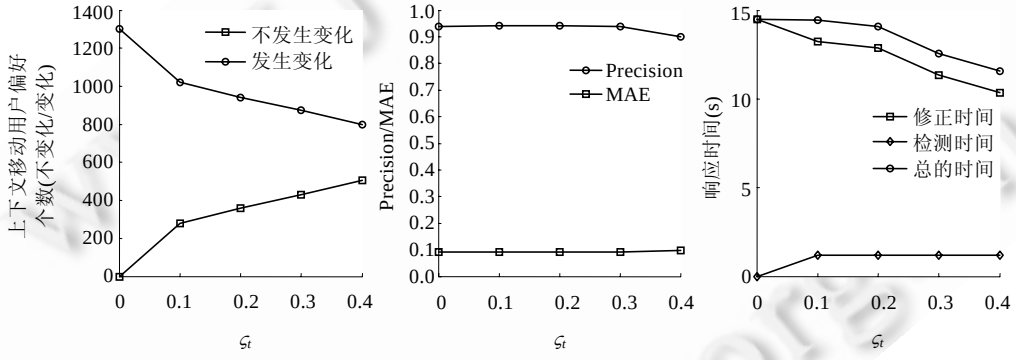
(a) 上下文移动用户偏好个数 (b) 学习结果 (c) 运行时间

Fig.2 Comparison of contextual mobile user preferences with different volatility thresholds (MIT dataset)

图 2 波动率阈值取不同值时,上下文移动用户偏好学习结果比较(MIT 数据集)

2) 上下文移动用户行为变化阈值ς_t和ς的影响

图 3 是在 MIT 数据集下,当ς_t取不同值时,在时间上下文约束下采用本文所提出的方法学习得到的上下文移动用户偏好的结果.其中,$Vol_{threshold}$=0.05,LSSVM 的参数为γ=1500,σ=0.17.图 4 是在 MIT 数据集下,当ς取不同值时,采用本文所提出的方法学习得到的上下文移动用户偏好的结果.其中,$Vol_{threshold}$=0.05,LSSVM 的参数为γ=1000,σ=0.2.图 5 是在模拟数据集下,当ς取不同值时,采用本文所提出的方法学习得到的上下文移动用户偏好的结果.其中,$Vol_{threshold}$=0.05,LSSVM 的参数为γ=23000,σ=0.35.



(a) 检测结果 (b) 学习结果 (c) 运行时间

Fig.3 Learning results of contextual mobile user preferences with different values of ς_t (MIT dataset)

图 3 ς_t取不同值时,上下文移动用户偏好学习结果(MIT 数据集)

(1) 对比图 3(a)和图 4(a)可知,当上下文类型增多时,上下文移动用户偏好个数也随之增加.例如,当ς=ς_t=0时,图 4(a)比图 3(a)中多 2 196 个上下文移动用户偏好,但图 4(a)中的上下文移动用户偏好个数小于 3 906.设时间上下文下移动用户偏好的数目为 m_t,则在时间、位置两个上下文约束下,移动用户偏好个数应为 n_l×m_t,其中,n_l表示位置上下文实例个数.在 MIT 数据集中,m_t=1302,n_l=3,因此上下文移动用户偏好数目应为 n_l×m_t=3×1302=3906.但由于某些上下文约束下移动用户没有使用移动网络服务的行为,即数据存在一些稀疏性,因此实际上下文移动用户偏好的个数比理论值要少一些.

(2) 由图 3(b)、图 4(b)和图 5(b)可知:当ς_t和ς取值为 0.2 时,修正后的上下文移动用户偏好精确度最高;当ς_t和ς取值为 0.4 时,其精确度小于直接学习(ς_t=0,ς=0)得到的上下文移动用户偏好的精确度.这是因为:

① 当ς_t和ς取值较小时,对上下文移动用户行为的变化范围比较严格,误判的几率比较小,因此不发生变化的上下文移动用户偏好学习后的准确率比较高;当ς_t和ς取值较大时,上下文移动用户行为的变化范围比较宽松,

误判的几率比较大,因此不发生变化的上下文移动用户偏好学习后的准确率比较低;

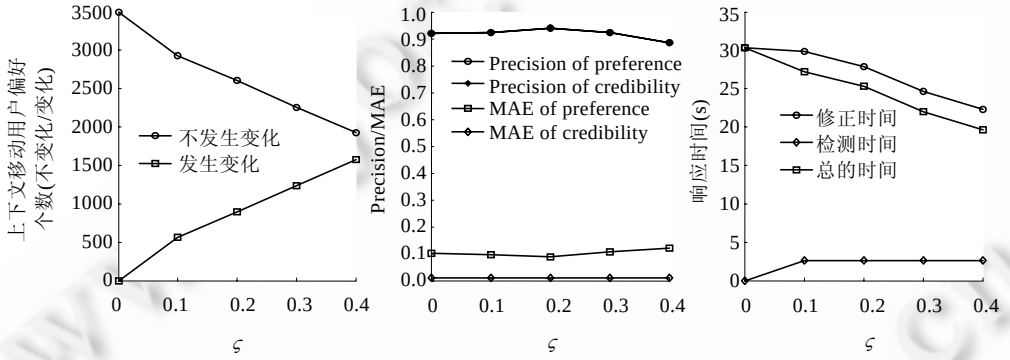
② 由于文中以新偏好值与已有偏好值相差小于 0.5 作为判定上下文移动用户偏好不发生变化的准则,所以,当公式(3)中的τ_1=1.75 时,只有在移动用户对某移动网络服务的变化量满足公式(24)的条件下才能判定上下文移动用户不发生变化.公式(24)表示如下:

$$abs(\log_{1.75}^{V_{inew}} - \log_{1.75}^{V_{iold}}) < 0.5 \tag{24}$$

其中,V_{iold} 表示移动用户对移动网络服务的原有使用量,V_{inew} 表示移动用户对移动网络服务新的使用量.由公式(24)推导可得:当$(V_{iold}-V_{inew})<0.2441\times V_{iold}$ 或$(V_{inew}-V_{iold})<0.3229\times V_{iold}$ 时,上下文移动用户偏好不发生变化.当ς=0.2 时,误判的几率为 0;当ς=0.3 且 $0.24\times V_{iold}<(V_{iold}-V_{inew})<0.3\times V_{iold}$ 时,会发生误判;当ς=0.4 且 $0.24\times V_{iold}<(V_{iold}-V_{inew})<0.4\times V_{iold}$ 或 $0.3229\times V_{iold}<(V_{inew}-V_{iold})<0.4\times V_{iold}$ 时,会发生误判;

③ 用 $P_{unchange}$ 表示不发生变化的上下文移动用户偏好学习后的精确度,P_{change} 表示发生变化的上下文移动用户偏好学习后的精确度.当ς_t 和ς取值较小时,误判的几率较小,$P_{unchange}$ 较大;随着ς_t 和ς的增大,$P_{unchange}$ 减小,但判定为不发生变化的上下文移动用户偏好的个数增加.当ς_t 和ς增加时,P_{change} 变化不大,且当$\varsigma_t\leqslant 0.3$,$\varsigma\leqslant 0.3$ 时,$P_{change}<P_{unchange}$;当ς_t=0.4,ς=0.4 时,$P_{unchange}<P_{change}$.

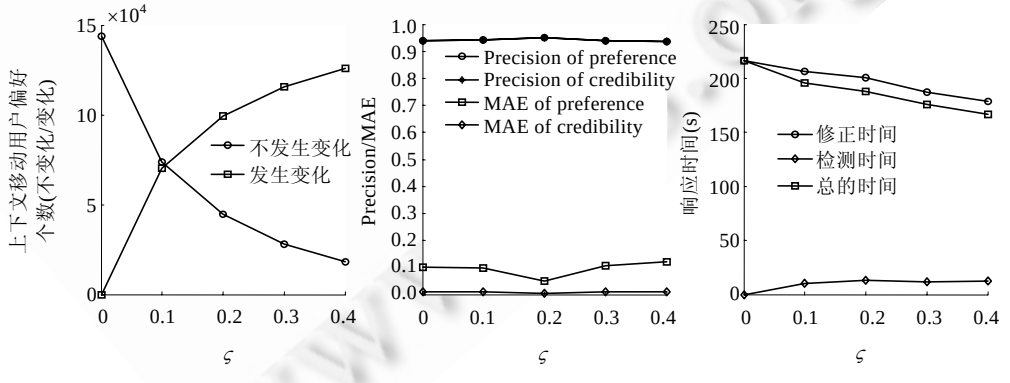
综合情形②和情形③可知,随着ς_t 和ς的增大,总的上下文移动用户偏好的精确度先增大后减小.因此,当ς_t=0.2,ς=0.2 时,学习后得到的精确度最高;当ς_t=0.4,ς=0.4 时,学习后得到的精确度小于直接提取时得到的精确度.



(a) 检测结果 (b) 学习结果 (c) 运行时间

Fig.4 Learning results of contextual mobile user preferences with different values of ς (MIT dataset)

图 4 ς取不同值时,上下文移动用户偏好学习结果(MIT 数据集)



(a) 检测结果 (b) 学习结果 (c) 运行时间

Fig.5 Learning results of contextual mobile user preferences with different values of ς (simulated dataset)

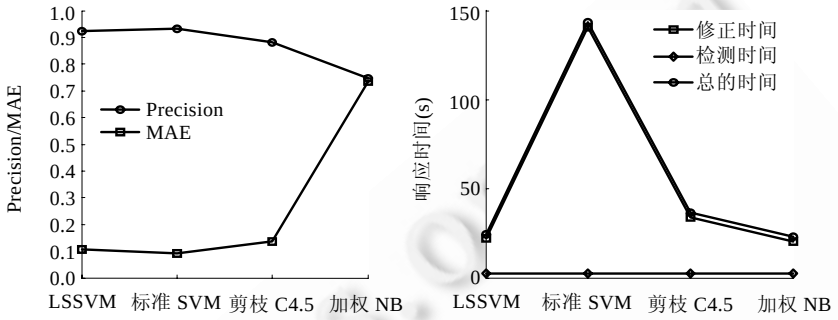
图 5 ς取不同值时,上下文移动用户偏好学习结果(模拟数据集)

(3) 由图 3(c)、图 4(c)和图 5(c)可知,当$\varsigma_t \geqslant 0.1$ 和$\varsigma \geqslant 0.1$ 时,其总的运行时间均比直接学习($\varsigma_t=0$ 和$\varsigma=0$)的运行时间要少.这是因为,当ς_t 或ς取上述值时,不发生变化的上下文移动用户偏好个数大于上下文移动用户偏好总数的 6.25%.例如:当$\varsigma_t=0.1$ 时,不发生变化的上下文移动用户偏好个数占总的上下文移动用户偏好个数的 21.51%;当$\varsigma=0.1$ 时,MIT 数据集中,相应的值为 16.21%,模拟数据集中为 48.74%.

通过以上分析并综合精确度和响应时间两方面的因素,本文选取$\varsigma_t=0.3$,$\varsigma=0.3$.

3) 与其他方法比较

图 6 为上下文移动用户行为变化时,通过不同分类方法学习得到的上下文移动用户偏好精确度和响应时间比较,其中,$\varsigma_t=0.3$,$\varsigma=0.3$,$Vol_{threshold}=0.05$,LSSVM 的参数$\gamma=1000$,$\sigma=0.2$.



(a) 学习结果 (b) 运行时间

Fig.6 Comparison of contextual mobile user preferences obtained by different classification methods (MIT dataset)

图 6 不同分类方法学习得到的上下文移动用户偏好比较(MIT 数据集)

由图 6 可知,LSSVM 的分类精确度比标准 SVM 分类方法的精确度差 0.95%,但其学习时间比标准 SVM 的学习时间少很多;与基于剪枝算法的 C4.5 分类方法相比,LSSVM 的分类精确度比其高 3.94%,学习时间快 50%;LSSVM 的学习时间比加权 NB 的学习时间长 10%,但其分类精确度比加权 NB 高 17.61%.由于本文实验数据比较稀疏,NB 分类受数据分布影响较大,因此其分类精确度较低,不适合移动网络中上下文移动用户偏好的学习.由于移动网络中移动用户对个性化移动网络服务的准确性及实时性提出了更高的要求,与基准方法相比,上下文最小二乘支持向量机分类学习方法更适合移动网络中上下文移动用户偏好的学习.

5 结束语

本文针对现有方法无法自适应地获取实时、准确的上下文移动用户偏好的问题,提出了一种基于上下文计算的移动用户偏好自适应学习方法.该方法通过分析移动网络中移动用户的日志行为,检测移动用户行为是否受上下文影响以及上下文移动用户偏好是否发生变化;然后,通过上下文最小二乘支持向量对变化的上下文移动用户偏好进行学习.由于本文采用先检测后更新的方法,并且在对变化的上下文移动用户偏好进行学习时本文采用了上下文最小二乘支持向量机方法,因此,与现有方法相比,本文提出的方法不仅保证了自适应学习的精确度,而且减小了自适应学习的响应时间.最后,通过仿真实验验证了本文所提出的方法在精确度、实时性方面均优于已有方法.

本文的不足之处是,在将上下文引入到最小二乘支持向量时,需要对上下文实例进行量化.现有的方法是给每个上下文实例一个数据,在进行分类时通过比较特征向量的距离进行分类,因此上下文的相似性是通过计算量化的值获取,失去了上下文原有的语义信息.未来将重点研究上下文的量化准则,并将信任度作为一种上下文引入,以期更准确地计算上下文之间的相似性,获取更准确的上下文移动用户偏好.

References:

[1] Do TMT, Gatica-Perez D. By their apps you shall understand them: Mining large-scale patterns of mobile phone usage. In: Proc. of the 9th Int'l Conf. on Mobile and Ubiquitous Multimedia. New York: ACM Press, 2010. 1−27. [doi: 10.1145/1899475.1899502]

[2] Chiu PH, Kao GYM, Lo CC. Personalized blog content recommender system for mobile phone users. Int'l Journal of Human-Computer Studies, 2010,68(8):496−507. [doi: 10.1016/j.ijhcs.2010.03.005]

[3] Kwok R. Phoning in data. Nature, 2009,458(7241):959−961. [doi: 10.1038/458959a]

[4] Wang LC, Meng XW, Zhang YJ. Context-Aware recommender systems: A survey of the state-of-the-art and possible extensions. Journal of Software, 2012,23(1):1−20 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4100.htm [doi: 10.3724/SP.J.1001.2012.04100]

[5] Wang LC, Meng XW, Zhang YJ. A cognitive psychology-based approach to user preferences elicitation for mobile network service. Acta Electronica Sinica, 2011,39(11):2547−2553 (in Chinese with English abstract). [doi: 10.1360/jos172518]

[6] Zhang ZZ, Zhai YQ, Xing HC. Implementation of preference reasoning by manipulating logical chain. Journal of Software, 2006, 17(12):2518−2528 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/2518.htm [doi: 10.1360/jos172518]

[7] Xu JJ, Liao SY, Li QD. Combining empirical experimentation and modeling techniques: A design research approach for personalized mobile advertising applications. Decision Support Systems, 2007,44(3):710−724. [doi: 10.1016/j.dss.2007.10.002]

[8] Mahmoud QH, Al-Masri E, Wang ZX. Design and implementation of a smart system for personalization and accurate selection of mobile services. Requirements Engineering, 2007,12(4):221−230. [doi: 10.1007/s00766-007-0051-3]

[9] Yao XL, Shu HY. Study on value-added service in mobile telecom based on association rules. In: Proc. of the 10th ACIS Int'l Conf. on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing (SNPD 2009). Daegu: IEEE Computer Society, 2009. 116−119. [doi: 10.1109/SNPD.2009.38]

[10] Hong JY, Suh EH, Kim JY, Kim SY. Context-Aware system for proactive personalized service based on context history. Expert Systems with Applications, 2009,36(4):7448−7457. [doi: 10.1016/j.eswa.2008.09.002]

[11] McBurney S, Papadopoulou E, Taylor N, Williams H. Adapting pervasive environments through machine learning and dynamic personalization. In: Proc. of the Int'l Symp. on Parallel and Distributed Processing with Applications (ISPA 2008). Sydney: IEEE Computer Society, 2008. 395−402. [doi: 10.1109/ISPA.2008.63]

[12] Xie HT, Meng XW. A personalized information service model adapting to user requirement evolution. Acta Electronica Sinica, 2011,39(3):643−648 (in Chinese with English abstract).

[13] Shi XW, Hua J. An adaptive preference learning method for future personalized TV. In: Proc. of Int'l Conf. on Integration of Knowledge Intensive Multi-Agent Systems. Milan: IEEE Computer Society, 2005. 260−264. [doi: 10.1109/KIMAS.2005.1427091]

[14] Suykens JAK, Vandewale J. Least squares support vector machine classifiers. Neural Processing Letters, 1999,9(3):293−300. [doi: 10.1023/A:1018628609742]

[15] Su JS, Zhang BF, Xu X. Advances in machine learning based text categorization. Journal of Software, 2006,17(9):1848−1859 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/17/1848.htm [doi: 10.1360/jos171848]

[16] Lin KM, Lin CJ. A study on reduced support vector machines. IEEE Trans. on Neural Networks, 2003,14(6):1449−1469. [doi: 10.1109/TNN.2003.820828]

[17] Stefanidis K, Pitoura E, Vassiliadis P. Modeling and storing context-aware preferences. In: Proc. of the 10th East European Conf. on Advances in Databases and Information Systems. Thessaloniki: Springer-Verlag, 2006. 124−140. [doi: 10.1007/11827252_12]

[18] Jiang JQ. Study on least squares support vector machine and its applications [Ph.D. Thesis]. Changchun: Jilin University, 2007 (in Chinese with English abstract).

[19] Eagle N, Pentland A, Lazer D. Inferring friendship network structure by using mobile phone data. Proc. of the National Academy of Sciences (PNAS), 2009,106(36):15274−15278. [doi: 10.1073/pnas.0900282106]

[20] Shi YC, Meng XW, Wang LC. A heuristic approach to identifying the specific household member for a given rating. In: Proc. of the 2nd Challenge on Context-Aware Movie Recommendation. New York: Association for Computing Machinery, 2011. 47−52. [doi: 10.1145/2096112.2096121]

附中文参考文献:

[4] 王立才,孟祥武,张玉洁.上下文感知推荐系统.软件学报,2012,23(1):1−20. http://www.jos.org.cn/1000-9825/4100.htm [doi: 10.3724/SP.J.1001.2012.04100]

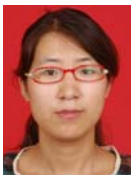
[5] 王立才,孟祥武,张玉洁.移动网络服务中基于认知心理学的用户偏好提取方法.电子学报,2011,39(11):2547−2553.

[6] 张志政,翟玉庆,邢汉承.偏好推理的逻辑链实现.软件学报,2006,17(12):2518−2528. http://www.jos.org.cn/1000-9825/17/2518.htm [doi: 10.1360/jos172518]

[12] 谢海涛,孟祥武.适应用户需求进化的个性化信息服务模型.电子学报,2011,39(3):643−648.

[15] 苏金树,张博锋,徐昕.基于机器学习的文本分类技术研究进展.软件学报,2006,17(9):1848−1859. http://www.jos.org.cn/1000-9825/17/1848.htm [doi: 10.1360/jos171848]

[18] 姜静清.最小二乘支持向量机算法及应用研究[博士学位论文].长春:吉林大学,2007.



史艳翠(1982－),女,河北保定人,博士生,主要研究领域为用户偏好自适应技术,推荐系统,个性化服务.

张玉洁(1969－),女,讲师,主要研究领域为智能信息处理,通信软件,网络服务.

孟祥武(1966－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络服务,通信软件,人工智能.

王立才(1984－),男,博士生,主要研究领域为推荐系统,智能信息处理.