文章编号：2096-1618(2016)增-0007-04

基于用户行为序列模式的性别分析与预测

吴东华， 常 征， 何 嘉
(成都信息工程大学，四川 成都 610225)

摘要：为了分析手机用户访问站点的行为与用户的性别特征的关系，通过分析处理手机用户访问站点的行为的历史数据，将用户的行为数据转化为序列的形式计算用户之间的相似度，并使用改进的编辑距离的 k 近邻序列分类算法完成性别的预测。结果表明采用序列的形式对行为数据进行处理均可以提高男女性别预测的准确率和召回率，从而证明手机用户访问站点的行为具有性别的特征。

关 键 词：计算机应用技术；数据挖掘；行为序列；编辑距离；kNN；性别预测

随着移动互联网的普及和推广，手机用户数量和所产生的流量数据明显增加，并已涉及用户生活中的衣食住行。移动互联网已经在人们生活中扮演着重要的角色，随之在手机上也产生大量的用户行为数据。近几年对用户行为数据的研究包括确定性别、年龄和地理位置等主题。在不同的领域都可以对用户行为数据进行挖掘分析，如用户的消费习惯、转化率和增长率都可以在电商、广告和金融等行业创造巨大的价值。传统的方法大多采用特征工程提取行为特征对用户建立用户画像预测性别，因此采用传统方法对原始数据进行特征抽取处理时会缺失数据本身的属性。因此引出采用序列的形式对原始行为数据进行处理。

序列的研究在很多领域都有实际的应用。序列分析作为数据挖掘中的一种重要方法，目的是在从大量序列数据中发现潜在的有用模式及信息，它在生物医学、金融行业、自然灾害预测等方面都有重要的作用[1]。在基因方面的研究发现，从现有的类别中对蛋白质序列进行分类可被用来发现学习新的蛋白质[1]；在行为异常检测方面，用户的系统访问活动时间序列常被用来检测是否有异常行为[2]；在信息检索方面，序列分类常用在将文档分类成不同的主题领域[3]。

由于序列数据的时序性和长度可变的情况，传统向量度量的分类算法已不能很好解决序列分类的问题。Keogh 等[4-5]介绍了当使用最近邻算法对等长的时间序列分类时，相比其他复杂的相似度计算方法用欧氏距离可以得到很高的准确率；DTW(动态时间弯曲距离)[6]的提出可以有效解决时间序列数据不等长的问题；但是针对大规模数据集时，DTW 在计算序列的距离时有很高的时间复杂度[7]。因此，将序列的研究方法应用在手机用户访问的站点上，并提出一种基于改进的编辑距离的 k 近邻序列分类算法预测手机用户的性别。

1 用户行为序列特征

1.1 用户行为序列分析

建立用户行为序列模式目的是描述手机用户访问站点的行为，通过行为序列模式反映用户访问站点的行为特征。研究表明大多用户模型都是从用户的兴趣和行为为出发点，在不同的领域两类模型所起到的作用也不同。比如在构建用户的用户画像时，为用户建立的兴趣模型计算用户的喜好并决定向用户推送何种服务，而行为模型则决定了用何种方式为用户推送服务。文献调研结果显示[14]，为用户构建知识图谱的模型大多从用户的兴趣为出发点提取特征构建模型，而针对用户的行为模型较少。研究的重点是将手机用户访问站点的历史记录转化为基于时间的用户行为序列模式，将提取不同性别手机用户访问站点的历史行为序列进行分析。

1.2 用户行为序列模式

在一定时期内用户的行为反映出的兴趣和习惯一般是比较固定的，而同性别的用户又具有相似的行为序列，比如，大多男性用户经常在一段时间内访问体育或游戏类的站点，女性用户经常访问社交类站点晒图。手机用户的这些行为反映了他们的习惯和行为规律，并产生较为明显的序列特征；而特征又可以反映出用户的性别，因此性别对用户的行为构建序列模式进行挖掘很有意义。文中定义序列为不同性别手机用户访

收稿日期：2015-07-30
基金项目：成都信息工程大学引进人才启动科研资助项目(KYTZ 201323)

问站点的行为。通过分析和挖掘用户的行为序列,可以有效地预测用户的性别,从而为用户提供个性化定制服务和推荐。

定义 1 序列。一个序列是有序,列表的事件。行为序列即为有时间顺序的连续事件、字符或者复杂的数据类型均可以代表一个事件,用户行为序列可表示为按照时间顺序排列的字符序列,一个行为序列中字符个数总和称为该序列的长度。

$S=\{S_1,S_2,\cdots,S_n\}$,

其中 $S_1,S_2,\cdots,S_n$ 为用户访问站点的类型。

2 用户行为序列性别预测算法

2.1 用户行为序列的相似度度量

对生物序列、行为序列等字符类型序列分类的参考文献还相对较少,Pavel P. Kuksa[8] 通过研究多元变量核函数的机器学习方法对生物基因序列进行分类。与时间序列不同,字符序列数据强调数据的类型属性及次序关系,而对时序数据的数值定量属性没有要求。因此在对字符序列和时间序列数据进行分类时的分析方法和度量方式也不同,文献[9]提出了根据用户行为浏览模式的相似度进行聚类的研究。序列编辑距离[9-10]不仅可以反映出用户行为在序列模式上的构成之间的差异,还能够区分用户行为时间上的差别。本文采用改进的编辑距离作为计算用户行为序列的相似性度量方法。

定义 2 用户行为序列 S 与 Q 之间的相似度

$$\mathrm{Sim}(S,Q)=(\mathrm{Sum}(S,Q)-\mathrm{Ldist}(S,Q))/\mathrm{Sum}(S,Q) \tag{1}$$

其中:$\mathrm{Sim}(S,Q)$ 为用户行为序列 S 与 Q 之间的相似度,$\mathrm{Sum}(S,Q)$ 表示序列 S 和 Q 字符串的长度之和,$\mathrm{Ldist}(S,Q)$ 是类编辑距离。将文献[9]中的 3 种字符操作赋予不同的权值,字符替换操作添加为 2,其他删除和插入操作不变,目的是为了细化手机用户在访问移动站点时所产生的行为序列。

针对手机用户访问移动站点的行为,在相同一段时间内 2 个用户分别访问的站点,$a,b,c,\cdots,z$ 分别代表他们所访问移动站点的类别,其行为序列可以用图 1的字符表示。两人在访问不同移动站点行为上的相同模式序列即等价于两个字符串序列的公共字串,而字符序列相似性在一定程度上也反映了手机用户在访问移动站点行为中的相似程度。

```
"860071020818639 ","a,c,e,b","0"
"860071021009337 ","a,c,b","0"
```

图 1 用户行为序列及性别

2.2 用户行为序列模式 k 近邻算法

文中采用 k 近邻算法对用户行为序列模式数据进行分类,使用定义 2 计算用户行为序列之间的相似度,具体的算法描述:

输入:初始化得到用户的行为序列 $S=\{S_1,S_2,\cdots,S_n\}$,训练数据集 $T=\{(S_1,y_1),(S_2,y_2),\cdots,(S_n,y_n)\}$

其中 $S_i\in S,y_i\in Y=\{0,1\}$ 为序列的类别,这里为用户的性别特征,$i=1,2,\cdots,N$

输出:用户行为序列 S 所属的类 y

(1)根据定义 2 给定的相似性度量方法,在训练数据集 T 中找出与 s 最邻近的 k 个点,涵盖这 k 个序列的 s 的邻域记住 $N_k(S)$;

(2) 在 $N_k(S)$ 中根据相似度的大小,对近邻的投票进行加权,相似度越大则权重越大(权重为距离平方的倒数);

(3) 通常 k 值采用交叉检验方法来确定(以男女性别预测的准确率和召回率为目标,$k=20$ 为基准递减测试)。

2.3 用户行为序列模式评价指标

在推荐系统、机器学习和数据挖掘等领域,精确率和召回率[11]是作为预测结果质量好坏的两个重要的评价指标。借鉴两个评价指标衡量用户行为序列模式预测男女性别的准确率。

精确率是评估预测的结果中目标成果所占的比例,其计算公式为

$$P=TP/TP+FP \tag{2}$$

召回率,就是从关注领域中,召回目标类别的比例,其计算公式为

$$R=TP/TP+FN \tag{3}$$

其中,TP:模型预测为正的正样本的数量;TN:模型预测为负的负样本的数量;FP:模型预测为正的负样本的数量;FN:模型预测为负的正样本的数量。

3 实验

3.1 数据说明

试验数据来源于某品牌手机的用户访问站点的历史记录。服务器端接收日志文件进行字段过滤,每天

大约 10 万条数据,人数共计 15270,男女比例约为 7∶3,过滤的字段维度主要有时间和用户每天在不同时间点访问站点的历史数据,并根据购买手机用户的性别对数据打标,来完成对缺失数据的预测和分析。

3.2　数据预处理

在海量数据存储平台的数据库中用户行为记录有数十亿条记录,传统的数据挖掘技术很难针对其进行处理,因此对数据采用随机抽样方法得到训练数据和测试数据。抽样采集到的手机用户行为数据大致将访问的站点分为 32 个类别,按照出现的频数将其排序,根据著名的 80/20 定律,前 21 个类别的站点占全部频数的 80%,因此将后 11 个类别不具有代表性的类别剔除,则按照用户经常访问站点的次数依次将 21 个种类按 a ~ u 字母顺序代表。在经过数据的加工和处理后,用户访问站点的历史转化为行为序列的模式,如表 1所示。

表 1　用户行为序列模式

用户	用户行为序列	用户性别
860071022155964	< k,n,d,a,m,c,g,e,l,h,b,f >	0
550861067487900	< p,a,c,e,b >	0
860486021453808	<a,c,b,e,d,g,f>	1
…	…	…
860071027827278	< n,d,a,c,e,h,b >	0

表 1 中第一列为用户标识即用户手机的 IMEI 号,第二列按用户行为在一段时间内顺序生成的行为序列,第三列为用户性别标识,“0”代表男性用户,“1”代表女性用户。

3.3　实验结果

为了验证对用户行为序列预测性别算法的有效性,对采用随机抽样方法得到的数据集采用 kNN 算法进行分类预测,距离计算分别使用最长公共子序列[12],编辑距离和文中改进的相似度度量。随机抽取数据集中 4/5 作为训练集,剩余的 1/5 作为测试集,并且保证每次验证的训练集和测试集相通。采用交叉检验的方法迭代 20 次,并选取这 20 次实验结果的精确率和召回率的平均值作为该数据集的性别预测结果。

在实验过程中,kNN 算法中的 k 值分别取 20 以内的奇数(偶数在算法最后投票阶段结果易出现平等事件),表 2 中的数据为 3 种不同相似性度量的 kNN 算法在不同 k 值中具有最高的精确率和召回率,括号里的整数为对应 k 的取值。

表 2　实验中 3 种相似度度量的 kNN 算法精确率和召回率

		最长公共子序列	编辑距离	改进的编辑距离
精确率 P	男	76.30(13)	73.29(9)	76.43(13)
	女	31.67(13)	26.07(9)	40.25(13)
召回率 R	男	55.62(13)	51.52(9)	87.31(13)
	女	47.61(13)	43.27.23(9)	52.39(13)

从表 2 可以看到,传统的最长公共子序列方法计算字符相似度在用户行为序列中要优于编辑距离处理行为序列,但是在编辑距离计算字符相似度基础上为 3 种操作赋予不同的权值,并引入类编辑距离可以去除两个完全不相同的行为序列其相似度却不为 0 的用户,从三者比较结果分析得出,改进的编辑距离的方法在对用户行为序列计算相似度可以有效提高男女性别预测的精确率和召回率。另一方面,kNN 算法在通过用户行为序列对性别分析和预测时准确率和召回率均超过 50%,能够有效地说明手机用户访问站点的行为与用户的性别特征属性的关系,而且能够进行较准确的预测。

4　结束语

提出一种将生物医学、金融行业中研究较为广泛的序列模式引入到手机用户行为中,并利用改进的编辑距离的 k 近邻序列分类算法,对手机用户行为序列进行分析从而预测用户的性别。经过对真实手机用户访问站点的数据集的实验分析,相比传统的基于公共子序列的分类算法,其在男女性别预测中的精确率和召回率都有显著的改善,从而验证了改进的编辑距离相似度的有效性。进一步的研究侧重方向拟考虑将用户行为序列按照不同时间段划分成子序列模式进行预测分析。

参考文献:

[1] M. Deshpande and G. Karypis. Evaluation of techniques for classifying biological sequences [A]. In PAKDD'02: Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining,2002:417 - 431.

[2] T Lane, C E Brodley. Temporal sequence learning and data reduction for anomaly detection[J]. ACM Trans. Inf. Syst. Secur. ,1999,2(3):295 - 331.

[3] 苏晨. 基于维基百科知识的文本分类技术研究[D]. 咸阳:西北农林科技大学, 2013.

[4] 徐璘俊. 智能分类算法在银行客户洗钱风险评

估中的应用研究[D]. 浙江:浙江大学, 2010.
[5] 王金龙. 全局和局部相结合的数据挖掘方法及应用研究[D]. 杭州:浙江大学, 2007.
[6] E. J. Keogh and M. J. Pazzani. Scaling up dynamic time warping for datamining applications. In KDD'00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining,2000:285-289.
[7] 孙磊. 健康管理中时序数据挖掘相关问题研究与应用[D]. 北京:清华大学, 2012.
[8] Pavel P Kuksa. Biological sequence analysis with multivariate string kernels[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013.
[9] 车高营, 张磊, 张禄旭. 基于序列模式的用户浏览行为提取与分析[J]. 计算机技术与发展, 2012, (9):9-12.
[10] Shasha Z D. Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems[J]. Siam Journal on Computing, 1989, 18(6):1245-1262.
[11] Campana S E. Accuracy,precision and quality control in age determination,including a review of the use and abuse of age validation methods[J]. Journal of Fish Biology,2001,59(2):197-242.
[12] 王映龙, 杨炳儒, 宋泽锋,等. 基因序列相似程度的LCS算法研究[J]. 计算机工程与应用, 2007, 43(31):45-47.
[13] 罗俊勤. 大众行为下社会网络的服务推荐研究[D]. 广州:华南理工大学, 2012.
[14] 刘春, 梁光磊, 谭国平. 基于用户兴趣变化融合的个性化推荐模型[J]. 计算机工程与设计, 2013, 34(8):2944-2950.
[15] 贺露. 基于社交网络的用户性格与行为分析[D]. 北京:北京邮电大学, 2014.