doi: 10.3969/j.issn.0490-6756.2017.06.012

基于集成分类器的用户属性预测研究

王斯盾[1], 琚生根[2], 周　刚[2], 刘玉娇[2]
(1. 后勤工程学院后勤信息与军事物流工程系, 重庆 401311; 2. 四川大学计算机学院, 成都 610065)

摘　要: 用户属性在个性化服务中具有重要的作用,利用手机数据进行用户属性预测逐渐成为新方向.利用手机应用类别均使用时长和应用类别个数,提出了基本属性与辅助属性的概念.首先对所有未标注样本的辅助属性离散化,将辅助属性基于类别的海灵格距离作为基本属性的特征权重,将基本属性与权重的乘积作为特征训练集成分类器中的各个基分类器,并引入随机森林中的带外样本准确率作为基分类器的权重,得到最终的分类结果.实验结果表明,本文所给出的集成分类器框架能够提高用户属性预测的效果.

关键词: 用户属性预测; 智能手机; 离散化; 海灵格距离; 特征权重
中图分类号: TP391　　**文献标识码**: A　　**文章编号**: 0490-6756(2017)06-1195-07

Research on demographic prediction based on ensemble classifiers

WANG Si-Dun[1], *JU Sheng-Gen*[2], *ZHOU Gang*[2], *LIU Yu-Jiao*[2]
(1. Department of Logistics Information & Logistics Engineering, Logistical Engineering University, Chongqing 401311, China; 2. College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: User attributes play an important role in personalized service. The prediction of the user's property based on mobile phone data has gradually become a new direction. In this paper, The authors use two independent attributes: average daily usage time and number of application categories. The basic attribute and the concept of the auxiliary attribute are proposed. In this paper, firstly, the auxiliary attributes of all unlabeled samples are discretized by non-supervised method. And then calculate the Hellinger Distance of auxiliary property categories, which is the characteristic weight of the basic attribute. Input the basic attributes and the characteristic weight to the base classifier of the ensemble classifier training model, introducing random forest with out of sample accuracy as the base classifier weights, finally the authors get the final classification results. The experimental results show that the ensemble classifiers framework can improve the effect of user attribute prediction.

Keywords: User attribute prediction; Smartphones; Discretization; Hellinger Distance; Feature weight

1　引　言

当今,移动手机已然成为人们日常生活中不可或缺的一部分.目前,移动手机的使用量也越来越大,据估计,到 2017 年,使用量将达到全部移动设备市场的 50%[1].鉴于移动手机数量的快速增长,研究移动手机对心理、社交以及经济的影响变得越来越重要.智能手机允许第三方开发 App 为用户提供不同的服务,开发者将其设计实现好的 App 发布到应用市场,用户可以下载并安装这些 App.

收稿日期: 2017-05-23
基金项目: 国家自然科学基金(61332066,81373239)
作者简介: 王斯盾(1993－), 男, 四川渠县人, 硕士生, 研究方向为智能检测与智能控制. E-mail: danube.live@qq.com
通讯作者: 琚生根. E-Mail: jsg@scu.edu.cn

在日常生活中,人们使用各种不同的 App 来取得相应的服务,而安装在用户手机上的 App 可以在不需要用户许可的情况下获得安装列表[2] 以及 App 的使用情况[3]. 因此,可以通过用户的手机应用列表预测用户的属性,为其提供更好的个性化服务. 当前,针对采用手机上的数据对用户人口属性预测的研究相对还不成熟,特别是在如何不侵犯用户隐私的情况下预测用户的人口属性[4,5] 方面还有极大的研究空间.

文献[6]首次根据用户手机上的安装列表对用户属性进行预测,主要是利用流行的 App 描述作为特征,构建多个支持向量机(Support Vector Machine,SVM)分类器对用户进行属性预测,该方法存在数据稀疏问题,为了进一步描述用户特征,Suranga 在 2015 年添加包括应用是否出现、应用所属类别,应用自身的描述等特征[7]. 为了进一步表达用户的兴趣,Qin 提出了采用一定时间内的类别使用频率(使用即为 1,未使用即为 0)作为模型特征. 在该方法中,虽然采用一定时间内的类别使用频率作为特征,具有一定的动态性,并采用贝叶斯方法对预测结果进行了平滑和优化,提高了准确率[8]. 但是该特征仍然较为粗糙,未反应应用使用的时间长短,无法进一步反应用户兴趣,且在预测用户属性(如性别、年龄等)的过程中,都是基于同一个假设,即所有的特征对于预测结果的影响力相同,没有考虑特征的权重问题,即未考虑到不同特征于预测结果的影响力区别问题.

针对以上问题,本文提出了前台均使用时间作为以应用类别个数作为辅助属性,应用类别均使用时长作为基本属性的集成分类器. 本文首先采用非监督方式对辅助属性进行离散化划分[9,10];然后,计算辅助属性基于类别的 Hellinger 距离,以此距离作为基本属性的特征权重,随后利用基本属性与权重的乘积作为特征输入到集成分类器中的各个基分类器(多分类-Softmax 回归算法[11])中进行模型训练,得到各个基分类器;最后,引入带外样本的准确率作为基分类器的权重,综合各个基分类器的结果作为最终的用户属性预测结果.

2 算法设计

2.1 算法框架

本文提出的集成分类器算法框架中有五个基算法,基算法为 Softmax 回归模型. 算法框架如图 1 所示.

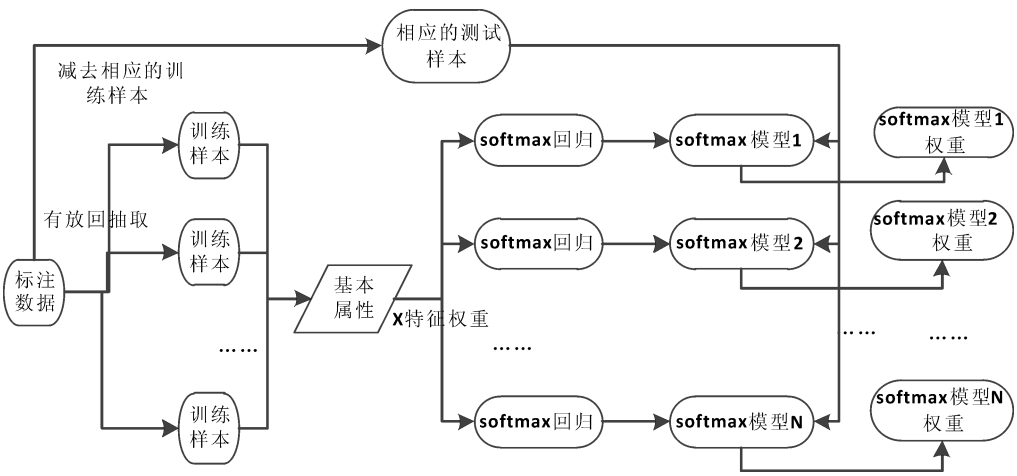

图 1 集成分类器算法框架图

Fig. 1 Framework of ensemble classifiers

首先从标注数据中有放回抽取训练样本,提取这些训练样本的特征表示,然后乘以相应的特征权重,随后输入到 Softmax 回归模型中进行模型训练,得到 Softmax 回归模型后,采用标注数据减去训练样本所得到的测试样本对模型进行测试,由此可以得到该模型的权重. 本算法具体步骤如算法 1 所示.

算法 1 基于辅助属性的集成分类器训练算法

输入 已标注数据 $L=\{(x_1^1,x_1^2,y_1),(x_2^1,x_2^2,y_2),\ldots,(x_N^1,x_N^2,y_N)\}$,未标注数据 $U=\{(x_1^1,x_1^2),(x_2^1,x_2^2),\ldots,(x_M^1,x_M^2)\}$(其中 $x^1\in\mathbf{R}^n$,表示基本属性,$x^2\in\mathbf{R}^n$,表示辅助属性).

输出 Softmax 回归模型 $f(f\in\mathbf{R}^N)$,模型权重 $W(W\in\mathbf{R}^N$,其中 $N=5)$.

1）利用以 U 及 L 获得特征权重 WA，$WA \in \mathbf{R}^n$；

2）从 L 中有放回的抽取训练样本；

3）根据上一步中抽取的训练样本，提取该样本的基本属性以及对应的类标；

4）计算基本属性与特征权重的乘积，使用该结果以及对应的类标对 Softmax 进行模型训练，根据代价函数；采用梯度下降法进行参数优化，得到 Softmax 回归模型；

5）利用 L 减去训练样本得到测试样本，对 Softmax 回归模型进行测试，得到 Softmax 回归模型在该样本上的准确率，该准确率作为对应 Softmax 回归模型的权重 W_i；

6）判断 Softmax 回归模型模型个数是否达到 N，若达到则结束算法，否则转入第 3）步继续执行；

7）最终输出回归模型以及对应的模型权重.

2.2　基本及辅助属性确定

本文所采用的两类属性分别为应用类别个数以及应用类别前台均使用时间[12]，应用类别为：体育竞速、健康健美、儿童教育、其他应用、其他游戏、出版物、办公效率、商务职场、图册漫画、地图出行、塔防策略、天气日历、女频、实用工具、影视视频、拍摄美图、新闻阅读、桌面美化、棋牌桌游、漫画、理财金融、生活购物、电子书、男频、益智休闲、短信通讯、移植汉化、系统优化、经营养成、网游、考试学习、聊天社交、角色扮演、跑酷动作、音乐节奏、音乐铃声以及飞行射击.

因本文只涉及应用类别个数以及应用类别前台均使用时间这两个属性，基本属性以及辅助属性的确定主要是根据属性对于模型类标预测的准确性进行判定，定义准确率高的属性为基本属性，另一个则为辅助属性[13].

2.3　离散化划分算法

在本文中，主要采用的是无监督聚类算法—K-means 算法对未标注的数据进行离散化的[14]. 对于聚类算法的评估通常采用 Davies-Bouldin 指数，简称 DBI. 它的作用是评估 K-means 算法中 k 值的取值. 分类个数的不同可以导致不同的值，值越小，分类效果越好[15].

2.4　特征权重计算

在特征权重计算过程中，首先离散化未标注的数据，随后根据结果离散化已标注数据，最后计算 Hellinger 距离，以 Hellinger 距离作为基本属性的特征权重. 基于辅助属性的特征权重计算步骤如算法 2 所示.

算法 2　基于辅助属性的特征权重计算

输入　已标注数据 $L=\{(x_1^1,x_1^2,y_1),(x_2^1,x_2^2,y_2),\ldots,(x_N^1,x_N^2,y_N)\}$，未标注数据 $U=\{(x_1^1,x_1^2),(x_2^1,x_2^2),\ldots,(x_M^1,x_M^2)\}$，其中 $x^1 \in \mathbf{R}^n$，表示基本属性，$x^2 \in \mathbf{R}^n$，表示辅助属性.

输出　特征权重 WA，$WA \in \mathbf{R}^n$.

1）离散化 U 中的辅助属性 x^2，即对 x^2 中每一维进行离散化处理，采用 K-means 算法进行数据的离散化；

2）根据第一步离散化结果，划分 L 中的 x^2，完成 L 中的 x^2 中的离散化；

3）计算 L 中离散化后的辅助属性与类标 y 之间的 Hellinger 距离，Hellinger 距离计算公式如下式所示.

$$d_H(X_1,\ldots X_K)=\prod_{m=1}^{M} d_H(X_i,X_j)$$

其中，$M=C_K^2$，i,j 为类标中的任意两类，且 $i<j$. $d_H(X_i,X_j)$ 的计算公式如下所示.

$$d_H(X_i,X_j)=\sqrt{\sum_{l=1}^{p}\left(\sqrt{\frac{|X_{il}|}{|X_i|}}-\sqrt{\frac{|X_{jl}|}{|X_j|}}\right)^2}$$

其中，X 表示属性中每个特征（即 App 所属类别）；$|X_i|$ 和 $|X_j|$ 分别表示数据集中类别为 i 样本和类别为 j 样本的个数；$|X_{il}|$ 和 $|X_{jl}|$ 表示特征 X 的值为 l 且分别属于用户属性类别 i 和类别为 j 的样本个数；p 表示属性 X 具有不同值的个数.

4）对所有的 Hellinger 距离采用 min-max 标准化进行归一化，归一化的结果即为辅助属性对应的特征权重，min-max 标准化也称为离差标准化，是对原始数据的线性变换，使结果值映射到[0-1]之间. 转换函数如下：

$$x^*=\frac{x-\min}{\max-\min}$$

其中，max 为样本数据的最大值；min 为样本数据的最小值.

5）最终输出特征权重.

2.5　集成分类器预测

模型训练完成之后，就可以对新来的未标注数据进行预测其类标信息. 具体的预测算法步骤如算法 3 所示.

算法 3　基于辅助属性的集成分类器预测算法

输入　Softmax 回归模型 f，$f \in \mathbf{R}^N$，模型权

重 W，$W \in \mathbf{R}^N$，未标注样本 x^1

输出 未标注样本 x^1 所属类标

1)基于 x^1，分别采用 f 计算该未标注样本属于所有类标的概率 $p(c_i \mid x^1)$，$i \in \{0,1\}$(性别预测)，$i \in \{1,2,3,4,5\}$(其他用户属性预测)；

2) 计算所有 $p(c_i \mid x^1)$ 与相应的模型权重乘积；

3) 判断上一步结果中的数值，计算所属不同类别的均值，最后获取最大均值的类标，将该类标赋予该样本，即

$$c = \arg\max_{c_k} \frac{1}{N}\{\sum_{j=1}^{N} p(c_i = c_k \mid x^1) \times W_j\}$$

其中，j 为 Softmax 回归模型；

4) 算法结束，返回类标 C.

3 实 验

3.1 实验数据

数据来源于联想公司. 具体的数据分布如表 1 和表 2 所示.

表 1 未标注数据

Tab. 1 Unlabeled data

不同用户数目	样本数目	涉及不同 App 个数
5348	17793	31229

表 2 有标注数据

Tab. 2 Labeled data

分组	A	B	C	D	E
样本个数	5236	3329	\	\	\
样本个数	1824	6355	2943	2254	1228
样本个数	2645	4310	1425	2392	1358
样本个数	2256	5193	2913	3286	1627

本文将上述数据(未标注数据、有标注数据)按照 1∶1 的方式划分为研究实验数据与对比实验数据，其中研究实验数据为模型中属性确定以及参数确定使用，对比数据最终训练测试模型使用. 在使用时，需要对研究实验数据以及对比实验数据进行再次划分，划分为训练数据集和测试数据集.

3.2 算法实验

3.2.1 *基本及辅助属性确定实验* 基本属性以及辅助属性的确定主要是根据属性对于模型类标预测的准确性进行判定. 本文分别利用应用类别个数以及应用类别前台均使用时间对 Softmax 回归模型进行训练并测试. 实验数据为有标注数据样本划分出来的研究实验数据. Softmax 回归模型预测用户属性效果如图 2～图 5 所示.

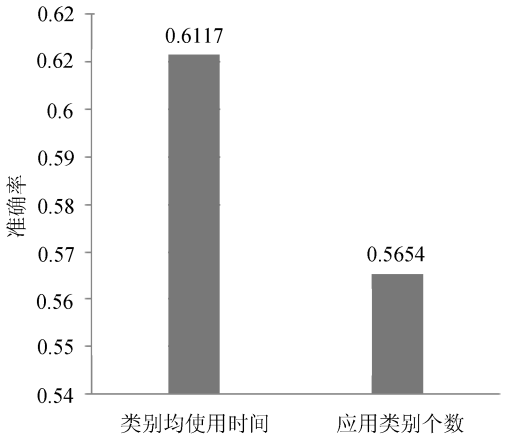

图 2 用户年龄预测图

Fig. 2 Prediction of user's age

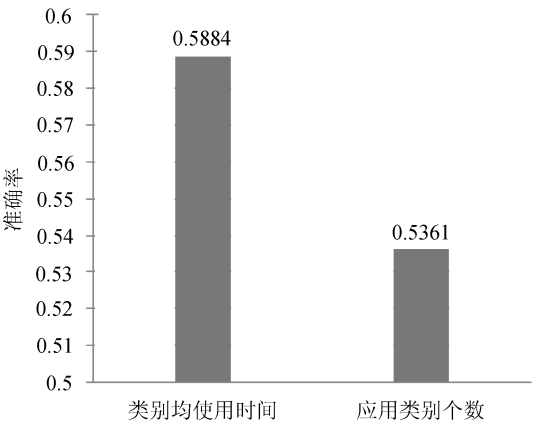

图 3 用户教育预测

Fig. 3 Prediction of user's degree

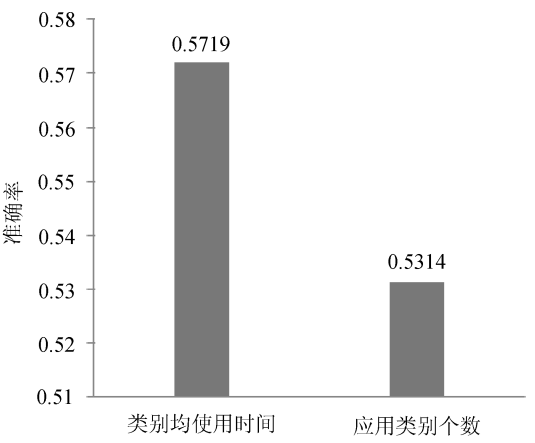

图 4 用户职业预测图

Fig. 4 Prediction of user's profession

由图 2～图 5 可以看出，应用类别前台均使用时间对模型进行训练预测效果比应用类别个数效果好，因此应用类别前台均使用时间作为基本属

性，应用类别个数作为辅助属性.

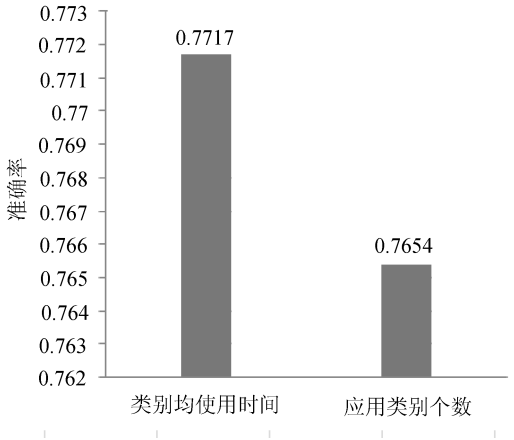

图 5　用户性别预测

Fig. 5　Prediction of user's sex

3.2.2　离散化实验　离散化实验数据涉及未标注数据，主要使用的属性为应用类别个数. 在离散化之前首先去除零数据，然后再进行离散化. 离散化的 DBI 值随着分类数目 K 的不同变化如图 6 所示.

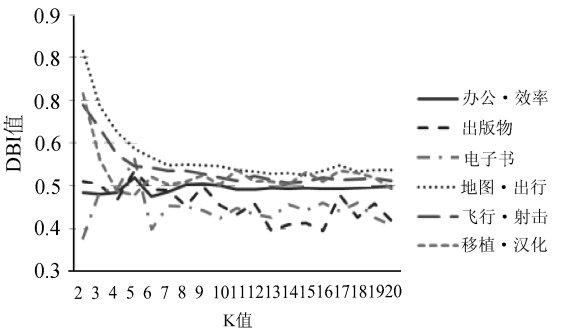

图 6　部分特征离散化 DBI 变化图 1

Fig. 6　DBI variation of partial feature discretization

图 6 是特征中办公・效率、出版物、电子书、地图・出行、飞行・射击、移植・汉化这几个维度的离散段数，即聚类数与 DBI 值的变化图示. 从图 6 可以看出，办公. 效率在聚类数目从 2～20 的变化中，最小值取值为 0.491，此时的 K 值为 12，因此在办公. 效率维度上进行离散化的时候，选取的离散段数为 12. 其他特征与之类似，K 值取 DBI 最小时的数目，由于篇幅限制，不再列出其他维度的聚类数与 DBI 值变化图，类推可以得出本文中 37 个维度的离散化分段数目，如表 3 所示.

在对未标注的数据聚类完成后，对标注的数据参照未标注数据的聚类结果进行分类离散化，最后对离散化的结果计算每个属性相对于类标的 Hellinger 距离，即为每个属性的特征权重.

表 3　应用类别属性离散分段表

Tab. 3　Discrete segment table of app-type

应用类别	体育竞速	健康健美	儿童教育	其他应用	其他游戏
离散段数	6	4	9	14	10
应用类别	出版物	办公效率	商务职场	图册漫画	地图出行
离散段数	13	12	10	5	15
应用类别	塔防策略	天气日历	女频	实用工具	影视视频
离散段数	10	8	14	9	12
应用类别	拍摄美图	新闻阅读	桌面美化	棋牌桌游	漫画
离散段数	8	6	10	9	5
应用类别	理财金融	生活购物	电子书	男频	益智休闲
离散段数	9	7	6	12	6
应用类别	短信通讯	移植汉化	系统优化	经营养成	网游
离散段数	6	5	8	12	6
应用类别	考试学习	聊天社交	角色扮演	跑酷动作	音乐节奏
离散段数	9	12	6	10	16
应用类别	音乐铃声	飞行射击			
离散段数	10	14			

3.2.3　基于集成分类器的用户属性预测实验　该实验通过将本文所提算法与前人的方法做对比实验，验证本文算法的优越性.

1）用户性别预测

文献[7,8]中都有对用户性别进行预测，因此这两篇文章的算法可以用来对比实验结果，与本节的基于辅助属性的集成分类算法进行对比.

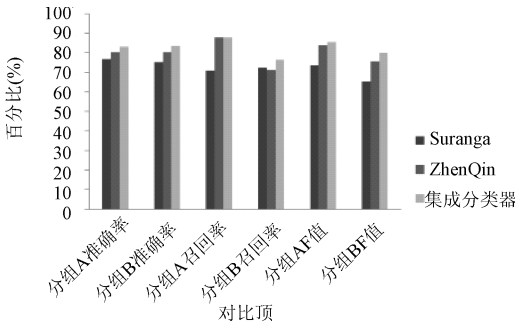

图 7　性别预测不同算法对比图

Fig. 7　Comparison of different algorithms for gender prediction

2）用户年龄预测

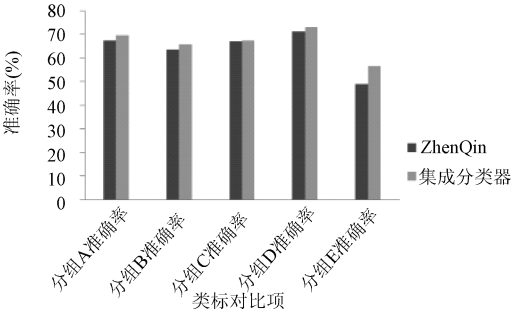

图 8　年龄预测不同算法准确率对比图

Fig. 8　The accuracy of different algorithms of age prediction

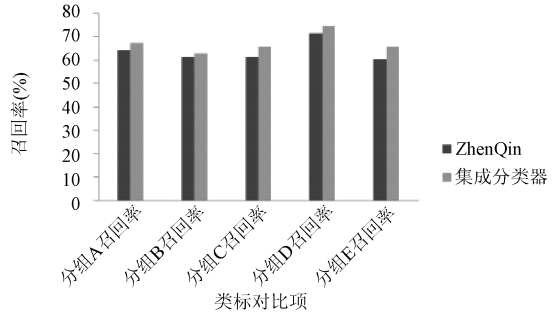

图 9 年龄预测不同算法召回率对比图

Fig. 9 The recall of different algorithms of age prediction

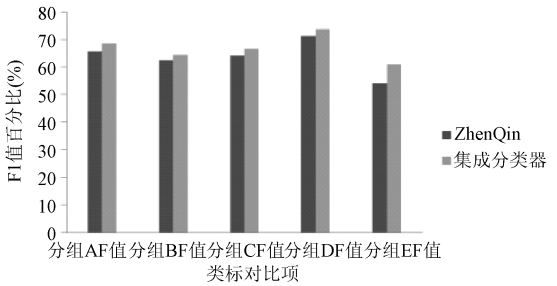

图 10 年龄预测不同算法 F_1 值对比图

Fig. 10 The F_1 of different algorithms of age prediction

3) 用户学历预测

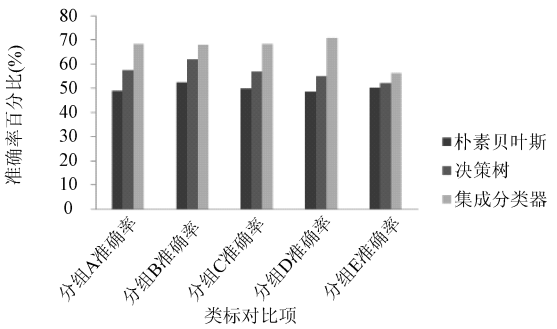

图 11 学历预测不同算法准确率对比图

Fig. 11 The accuracy of different algorithms of degree prediction

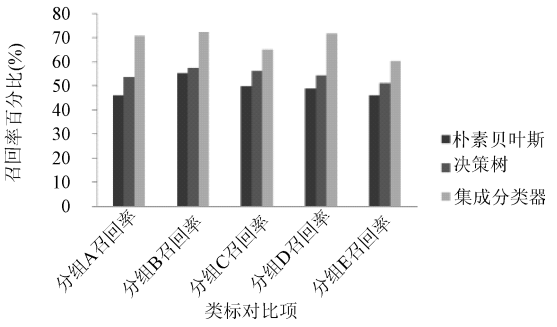

图 12 学历预测不同算法召回率对比图

Fig. 12 The recall of different algorithms of degree prediction

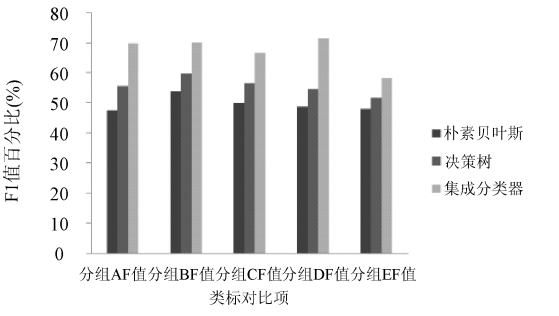

图 13 学历预测不同算法 F_1 值对比图

Fig. 13 The F_1 of different algorithms of degree prediction

4) 用户职业预测

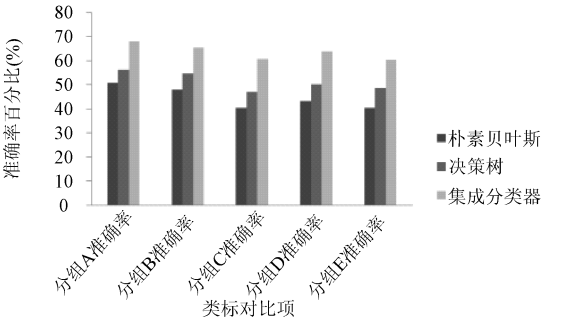

图 14 职业预测不同算法准确率对比图

Fig. 14 The accuracy of different algorithms of profession prediction

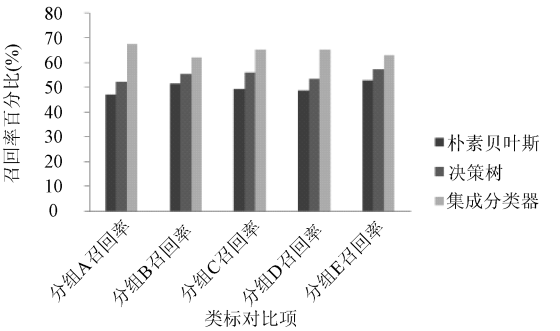

图 15 职业预测不同算法召回率对比图

Fig. 15 The recall of different algorithms of profession prediction

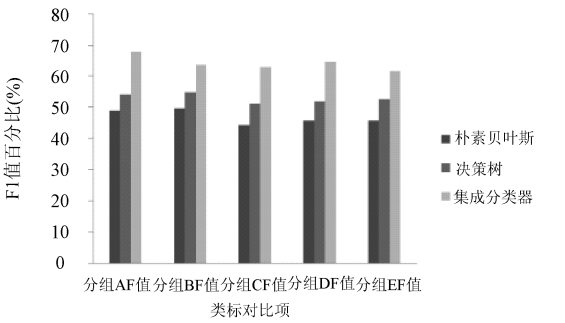

图 16 职业预测不同算法 F1 值对比图

Fig. 16 The F1 of different algorithms of profession prediction

由图 7～图 16 可以明显看出，本文所提出的基于辅助属性的集成分类算法在用户性别、年龄、教育、职业预测准确率、召回率以及 F1 值方面优于前人算法、朴素贝叶斯、决策树算法，说明了本文所提算法的有效性.

4　结　论

本文提出了以应用类别个数作为辅助属性，应用类别前台均使用时间为基本属性的集成分类器方法. 本文首先采用非监督方式对辅助属性进行离散化划分，然后计算辅助属性基于类别的 Hellinger 距离，并以此距离作为基本属性的特征权重，随后利用基本属性与权重的乘积作为特征输入到集成分类器中的各个基分类器（多分类-Softmax 回归算法）进行模型训练，并引入带外样本的准确率作为模型权重，最后综合各个基分类器的结果为最终的用户属性分类结果.

本文基本上完成了模型的主要任务，实现了算法的关键部分. 基于时间和实验条件的约束，还有很多地方需要改进. 如在考虑类不平衡的情况，只是简单地通过多采样方式来保持样本的类平衡，没有考虑通过其他方法消除类不平衡的影响. 另外，在本文中预测用户属性时采用的框架以及基算法都是一样的，可能在预测某一种用户属性的时候效果不如特定算法对用户属性预测效果好，且仅仅涉及到了两类特征，对于手机用户的兴趣把握可能还需要进一步的研究.

参考文献：

[1] eMarketer. Smartphone adoption tips past 50% in major markets worldwide[EB/OL]. (2013-05-29). [2017-03-20]. http://www.emarketer.com/Article/Smartphone-Adoption-Tips-Past-50-Major-Markets-Worldwide/1009923.

[2] Grace M C, Zhou W, Jiang X, *et al*. Unsafe exposure analysis of mobile in-App advertisements[C]// Proceedings of the 5th conference on Security and Privacy in Wireless and Mobile Networks. New York: ACM, 2012.

[3] ihasApp. iOS App detection [EB/OL]. (2016-01-01). [2017-03-20]. http://www.ihasApp.com.

[4] Malmi E, Weber I. You are what apps you use: demographic prediction based on user's apps[C]// Proceedings of the Tenth International AAAI Conference on Web and Social Media. (ICWSM 2016). Palo Alto: AAAI Press, 2016.

[5] Wang Y, Tang Y, Ma J, *et al*. Gender prediction based on data streams of smartphone applications[M]. Berlin: Springer, 2015.

[6] Seneviratne S, Seneviratne A, Mohapatra P, *et al*. Predicting user traits from a snapshot of Apps installed on a smartphone[J]. ACM Sigoood Mobile Comp Com, 2014, 18: 1.

[7] Seneviratne S, Seneviratne A, Mohapatra P, *et al*. Your installed Apps reveal your gender and more[J]. ACM Sigcomm Mobile Comp Com, 2015, 18: 55.

[8] Qin Z, Wang Y, Xia Y, *et al*. Demographic information prediction based on smartphone Application usage [C]//Proceedings of 2014 International Conference on Smart Computing. New York: IEEE, 2014.

[9] Mohamed M S, Hirani A N, Samtaney R. Discrete exterior calculus discretization of Incompressible Navier-Stokes equations on simplicial meshes[J]. Acad Press Prof, 2016, 312: 175.

[10] 刘汉清，朱敏，苏亚博，等. 一种考虑用户兴趣转移特征的协同预测模型[J]. 四川大学学报：自然科学版，2016，53：71.

[11] Ding Y, Ma J, Tian Y. Health assessment and fault classification for hydraulic pump based on LR and softmax regression[J]. J Vibroeng, 2015, 17: 1543.

[12] 孙煜哲，段磊. 基于共同主题群的用户影响力评估研究[J]. 四川大学学报：自然科学版，2014，51：81.

[13] 吴少华，马晓娟，胡勇. 基于改进 PageRank 算法的微博用户影响力评估[J]. 四川大学学报：自然科学版，2015，52：419.

[14] 冯波，郝文宁，陈刚，等. K-means 算法初始聚类中心选择的优化[J]. 计算机工程与应用，2013，49：182.

[15] Sina. 浅说 Davies-Bouldin 指数[EB/OL]. (2012-05-27). [2017-03-20]. http://blog.sina.com.cn/s/blog_65c8baf901016flh.html.