引用格式: 王乐,倪维健,林泽东,等. 基于模型堆叠的上网行为日志用户画像方法[J]. 山东科技大学学报(自然科学版),2018, 37(5):70-78.

WANG Le,NI Weijian,LIN Zedong,et al. Online behavior log data user portrait method based on model stacking [J]. Journal of Shandong University of Science and Technology (Natural Science), 2018, 37(5):70-79.

基于模型堆叠的上网行为日志用户画像方法

王　乐[1],倪维健[1],林泽东[1],曾庆田[2]

(1. 山东科技大学 计算机科学与工程学院, 山东 青岛 266590;2. 山东科技大学 电子通信与物理学院, 山东 青岛 266590)

摘　要: 上网行为日志数据中包含着大量的用户个性化信息,如何充分挖掘和分析这些信息至关重要。在分析上网行为日志数据的重要性后,提出了一种基于上网行为日志用户画像方法。在该方法中,首先通过特征选择和特征提取方式构建用户特征集,然后利用模型堆叠的技术组合多种单一分类器,构建用户画像模型。利用校园网行为日志数据对性别、年级、年龄三个维度进行用户画像,实验结果表明了所提方法的有效性。

关键词: 特征选择;特征提取; 用户画像; 模型堆叠

中图法分类号: TP391　　**文献标志码:** A　　**文章编号:** 1672-3767(2018)05-0070-09

DOI: 10. 16452/j. cnki. sdkjzk. 2018. 05. 010

Online Behavior Log Data User Portrait Method Based on Model Stacking

WANG Le[1], NI Weijian[1], LIN Zedong[1], ZENG Qingtian[2]

(1. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China; 2. College of Electronic Communication and Physics, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)

Abstract: It is significant to know how to fully tap and analyze users' personal information contained in their online behavior log data. After the investigation into the importance of online behavior log data, this paper proposed a online behavior log data user portrait method based on model stacking. In this method, user feature set was firstly constructed through feature selection and feature extraction, and then multiple single classifiers were combined by using stacking technology to construct user portrait model. This paper conducted user portrait based on campus network log behavior data in dimensions of gender, grade, and age. The experiment results has proved the effectiveness of the proposed method.

Key words: feature selection; feature extraction; user portrait; model stacking

收稿日期: 2017-12-12

基金项目: 国家自然科学基金项目(61472229,61602278,61702306,31671588);山东省科技发展计划项目(2016ZDJS02A11,ZR2017MF027);教育部人文社会科学研究项目(16YJCZH012,16YJCZH041 和 16YJCZH154);山东省泰山学者攀登计划专项和山东科技大学科研创新团队支持计划项目(2015TDJH102)

作者简介: 王　乐(1991—),女,山东德州人,硕士研究生,主要研究方向为数据挖掘,E-mail:wanglechp@163. com.
倪维健(1981—),男,山东临沂人,副教授,博士,主要研究方向为数据挖掘、机器学习.
曾庆田(1976—),男,山东高密人,教授,博士生导师,主要研究方面为流程挖掘、机器学习,本文通信作者.
E-mail:qtzeng@163. com

随着互联网技术的不断发展和普及,网民的数量迅速上升,根据中国互联网络信息中心(CNNIC)发布的《第 40 次中国互联网发展状况统计报告》,截止 2017 年 6 月,我国网民规模达 7.51 亿。网民在遨游网络的同时,在网络中留下了海量的上网行为日志数据。分析上网行为日志数据,挖掘出用户兴趣、喜好、基本属性(性别、年龄等),可以为个性化推荐、精准营销、商业决断分析、风险控制等提供数据支持。

用户画像作为大数据技术的重要应用之一,为分析和挖掘上网行为日志数据提供了可靠的方法。用户画像是由交互设计之父 Alan Cooper 提出的[1],定义用户画像为真实用户的虚拟代表,根据一系列用户的真实数据来挖掘出目标用户。用户画像根据用户的基本属性、生活习惯和上网行为等信息,筛选出一类用户标签,给用户信息进行结构化处理。其目的是在多维度上构建用户的标签属性,利用这些标签属性构造用户真实的特征,可用于描述用户的兴趣、偏好、特征等。Fawcett T 等[2]利用规则发现的方法在大量电话记录中发现欺诈行为标签,利用这些标签构建用户画像模型,该方法可以产生高可信度的报警。Adomavicius 等[3]展示了针对个性化的用户画像模型,利用分类规则、关联规则等数据挖掘方法来发现潜藏在用户商品交易记录中的行为档案信息。Nasraoui 等[4]根据动态网站的网络日志数据构建了动态可演化的用户行为画像模型,提出的网络使用日志挖掘框架可以挖掘、追踪和验证动态的多方面用户画像信息。陈志明等[5]基于“知乎”网站的数据,构建了基于用户基本属性、社交属性、兴趣属性和能力属性四个维度的动态用户画像模型,并通过对“知乎”网站 PM2.5 话题下 1303 位用户进行实证分析,得出的动态用户画像模型可以很好的区分用户的能力。Burger 等[6]通过提取 Twitter 的个人简介中隐藏的特征标签构建用户画像方法,利用 SVM[7]、朴素 Bayes[8]和 Balanced Winnow2[9]等分类器针对性别标签进行实验,得到了较好的实验效果。Iglesias 等[10]根据用户在 Unix Shell 上的命令日志数据研究用户画像,获得了计算机用户的行为画像。郭光明[11]基于微博行为数据进行了用户信用画像的研究,利用带有 L2 正则的逻辑斯蒂回归分类器对用户进行分类,实验结果表明学习出的用户行为模式可以很好地解释用户的信用标签。但由于上网行为日志数据的复杂性,传统的用户画像方法不能很好的应用于上网行为日志数据中。本文通过分析校园网日志的特点,提出了一种多维度标签用户画像方法。结合五种特征选择算法构建多指标融合的特征选择方法,融合二元特征和关联规则特征提取方法构建标签库,在两层叠加式框架中组合支持向量机、随机森林、决策树、朴素贝叶斯和逻辑斯蒂回归五种单一分类器模型构建基于 Stacking 的用户画像。实验结果证实了本文用户画像方法比单一分类模型在识别用户性别、年级、年龄属性的准确性上有较大提高。

1　基于上网行为日志的用户画像框架

参考数据挖掘的一般研究流程框架,本研究基于上网行为日志的用户画像框架如图 1 所示,主要包括特征工程和分类模型两个关键环节。其中,特征工程是通过特征选择和特征提取(基本单特征、二元特征、关联规则特征)来构建标签库;而分类模型是利用支持向量机、逻辑斯蒂回归、决策树、随机森林和朴素贝叶斯五种单一分类器模型构建 Stacking 组合模型。

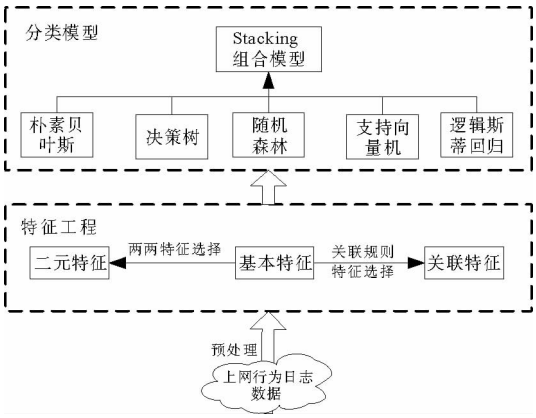

图 1　用户画像框架

Fig. 1　The framework of user portrait

2　用户特征选择与提取

2.1　构建标签库

标签是用户特征的符号标识。标签具有两个重要特征,一是具有一定的种群性,可以在一定程度上抽样出概括事务的特征;二是可以使用符号来表示用户的某一类特征,这个符号可以是中文、英文,也可以是数字。标签库则是对标签进行集中管理,最终用于对用户行为、属性的标记。

本研究基于上网行为日志数据的特性构建了三级标签。一级标签分为学生基本属性和行为标签两部分,二级标签则是对一级标签的细分,三级标签是标签库中最详细的标签描述。标签库的分级层结构如图 2

所示。

2.2 特征选择

特征选择[12]是模型识别的关键因素之一,特征选择结果的好坏直接影响分类结果的精度,因此需要有效的方法进行特征选择,提取对标签区分能力较高的特征,同时删除无用的噪声特征。

目前,特征选择的方法有很多,但是针对实际问题的研究仍存在各自不足。本文融合 Pearson 相关系数(皮尔森相关系数)[12]、Ridge Regression(岭回归)[13]、Chi(卡方检验)、RandomForest[14](随机森林)和 Stability-Selection(基于随机 lasso 的稳定性选择)[15]五种不同类型的特征选择算法构建多指标融合的特征选择方法,有效地避免了一种特征选择方法的不稳定性。本文特征选择方法的架构如图 3 所示。

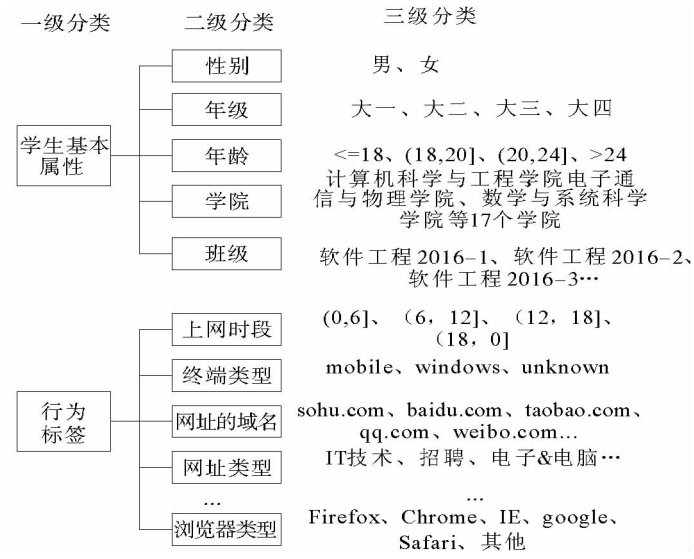

图 2 标签库分级结构图

Fig. 2 Hierarchical structure map based on label Library

假设 $\boldsymbol{X}=(X_1,X_2,\cdots,X_k)$ 为 n 维 k 列满秩矩阵,代表由 k 列不同属性特征组成的 n 维样本训练集合,其中 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{in})^{\mathrm{T}}$ 代表第 i $(i\leqslant k)$列训练样本。$\boldsymbol{Y}=(\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_l)$ 代表 n 维 l 列目标类向量(类标签),其中 $\boldsymbol{Y}_j=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$ 代表第 j $(j\leqslant l)$列类标签。$\{F_1,F_2,F_3,F_4,F_5\}$ 代表 Pearson、Ridge Regression、Chi、RandomForest、Stability-Selection 五种特征选择方法。$coef_j(j\in[1,5])$ 代表由第 j 类特征选择方法得到的 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的相关系数集合。

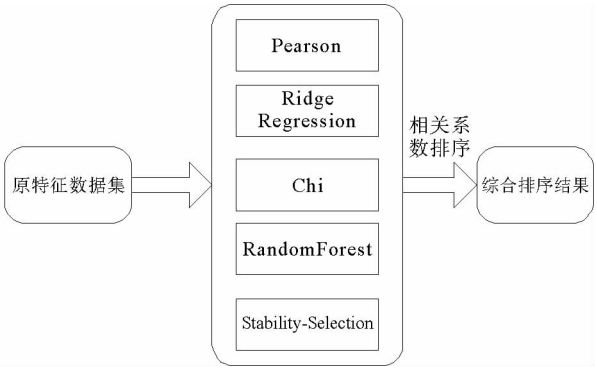

图 3 特征选择方法架构图

Fig. 3 Architecture diagram of feature selection method

由图 3 可以看出,本研究多指标融合的特征选择方法,首先利用五种特征选择算法得到 $\boldsymbol{X}$ 与 $\boldsymbol{Y}$ 的相关系数 $coef_j$,然后通过 k 种不同特征对相关系数进行排序,最后综合五种特征选择排序结果选择出排名前 t 的特征子集 F_select。

本研究的特征选择算法如图 1 所示。

2.3 特征提取

本研究提取了三类特征用于训练用户画像模型:基本特征、二元特征和关联特征。

1) 基本特征

首先对数据库数据属性进行筛选、离散化处理得到离散化特征;然后根据特征选择方法选择特征。假设离散化属性直接分为 K 个特征,例如性别直接划分为男、女两个特征。非离散化属性则采用等距离划分算法与等频率划分算法。

算法 1: 多指标融合的特征选择算法

输入:属性特征 $\boldsymbol{X}$,标签特征 $\boldsymbol{Y}$,特征选择方法 $Methods=\{F_1,F,_2F_3,F_4,F_5\}$。

输出:特征子集 F_Select。

步骤:

1. $F\_Select \leftarrow \phi$
2. $Merge \leftarrow \phi$// 相同特征的相关系数排序集合
3. **for each** $F_i \in Methods$ **do**
4.　$coef_i \leftarrow F_i(\boldsymbol{X},\boldsymbol{Y})$
5.　$Sort(coef_i)$
6.　$Merge = Merge \cup coef_i(\boldsymbol{X},\boldsymbol{Y})$
7. **end for**
8. $Sort(Merge)$// 综合 5 种特征选择结果排序
9. **for each** $i \in [1,t]$//**do** 选择排名前 t 的特征子集
10.　$F\_Select = F\_Select \cup Merge[i]$
11. **end for**
12. **return** F_Select // 输出特征子集

等距离划分算法:在每个属性上,根据给定的参数把属性值划分为距离相等的断点段,假设某个属性的最大属性值为 $\chi_{\max}$,最小属性值为 $\chi_{\min}$,用户给定的参数为 K,则断点间隔为 $\delta=(\chi_{\max}-\frac{\chi_{\min}}{K})$。为此得到此属性上的断点为 $\chi_{\min}+i\delta(i=0,1,\cdots,K)$。

等频率划分算法:根据给定的参数 K 把 m 个对象分段,每段中有 m/K 个对象。首先将此属性在所有实例上的取值排序,然后每隔 m/K 取一个值作为断点。

2) 二元特征

提取二元特征是指对不同类型的特征进行两两组合后利用选定的特征选择算法进行特征选择。

假设离散化之前的特征集合为 $F=\{f_1,f_2,\cdots,f_{m-1},f_m\}$,离散化之后的特征集合为 $F_1=\{f_{1\_1},f_{1\_2},\cdots,f_{k\_1},f_{k\_2},\cdots,f_{k\_e},\cdots,f_{m-1\_r},f_{m\_t}\}$,其中,$f_m$ 表示一共有 m 个特征,f_{k_e} 表示特征 f_k 离散化后的第 e 个离散特征。则组合后的二元特征集合为 $F_2=\{f_{1\_1}Xf_{2\_1},\cdots,f_{1\_1}Xf_{k\_e},\cdots,f_{m-1\_r}Xf_{m\_t}\}$。

3) 关联特征

利用关联规则挖掘算法计算特征组合与目标属性 $\boldsymbol{Y}$ 的关联强度,然后为每个特征组合计算在该目标属性中不同取值上的关联强度熵,据此得到该特征组合的权重 $weight$,根据所有特征权重集合 $results\_weight$ 排序得到排名前 p 的特征组合,用于后续的模型训练。本文提取关联特征的算法如算法 2 所示。

算法 2 中,L_K 代表频繁 K 项集,$all\_frequents$ 代表频繁项集集合,$apriori\text{-}gen(L_{K-1})$函数根据 L_{k-1} 中的频繁项集连接、剪枝产生候选 K 项集 C_K;D 是由基本特征 $\boldsymbol{X}$ 组成的数据集,函数 $Sort(results_weight)$对所有关联特征根据权重进行排序。

得到上述三种特征后,按照如下方式构建模型的特征库,其中 top_n 表示选取排名前 n 的特征。

$$(\text{基本特征})\times(top\_n\ \text{二元特征})\times(top\_m\ \text{关联特征}) \tag{6}$$

算法 2:关联特征提取算法

输入:属性特征 $\boldsymbol{X}$,标签特征 $\boldsymbol{Y}$,最小支持度计数阈值 min_sup。

输出:强关联规则特征子集。

步骤:

1. $L_1 = apriori\_gen(0)$// 生成频繁 1 项集
2. $all\_frequents = L_1$
3. $results\_weight = \phi$
4. **for**($k = 2; L_{K-1} \neq \phi, K++$)**do**
5. $C_K = apriori - gen(L_{K-1})$// 得到候选 K 项集
6. **forall** transaction t ∈ D **do**
7. $C_t = subset(C_K, t)$// 得到事务 t 中包含的候选集
8. **for each** candidate c ∈ C_t **do**
9. $c.count++$
10. **end for**
11. **end for**
12. $L_k = \{c \in C \mid c.count \geqslant \text{min\_sup}\}$
13. $all\_frequents = all\_frequents \cup L_K$
14. **end for**
15. **for each** $item_i \in all\_frequents$ **do**
16. $weight_i = -\sum_{j=1}^{m} p(y_j \mid Item)\log p(y_j \mid Item_i)$
17. $results\_weight = results\_weight \cup weight$
18. **end for**
19. $Sort(results\_weight)$// 对关联特征进行排序
20. ***return*** $To\_p(results\_weight)$// 输出关联特征子集

3 Stacking 组合模型

Stacking 组合模型是指将多种分类器组合在一起来取得更好表现的一种集成学习模型。本文采用逻辑斯蒂回归(logistic regression,LR)、支持向量机(support vector machine,SVM),决策树(decision tree,DT),随机森林(random forest,RF)和 Bayes(朴素贝叶斯)五种单一分类器模型构建 Stacking 组合模型,采用两层叠加式框架,第一层对数据集 data 进行 K 折交叉验证训练多个单一分类器模型,然后将第一层训练模型的输出加入原训练样本集作为输入,在元分类器下训练第二层模型,得到一个最终输出。其中单一分类器模型中预测效

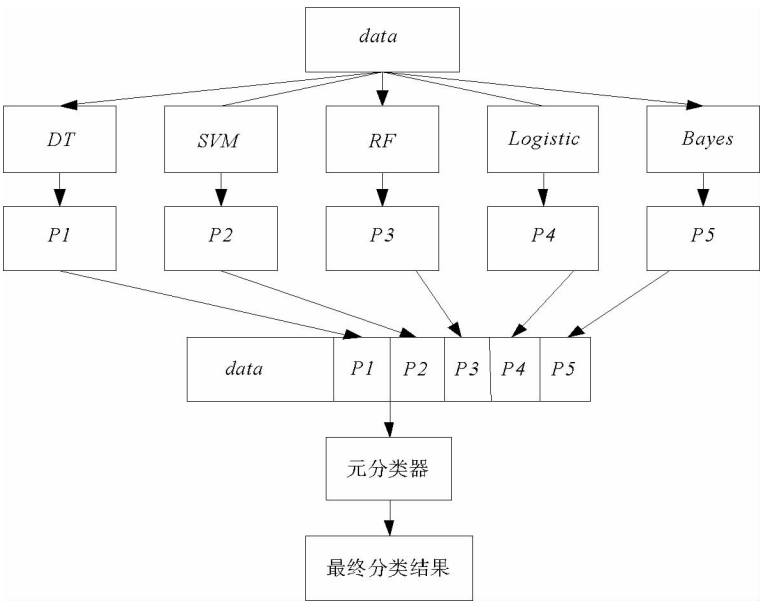

图 4 基于 Stacking 的组合模型架构

Fig. 4 Architecture diagram of stacking model

果最好的模型作为第二层模型训练的元分类器。基于 Stacking 的组合模型架构如图 4 所示。

Sacking 算法如算法 3 所示。

算法 3:Stacking 模型组合算法
输入:数据集 data,特征选择方法 Models,交叉验证次数 K。
输出:分类结果 *output*
步骤:

```
output = ϕ
foreach clf ∈ Models do
    predicted = ϕ// 得到候选 K 项集
    foreath (train,test) ∈ cross_validation(K,data)do
        clf.fit(train)// 训练分类模型
        predicted = predicted ∪ clf.predict(test)// 得到预测结果
    end for
    data = data ∪ predicted// 将预测结果加入预测样本特征
end for
output = bestmodel.fit(data)// 对关联特征进行排序
return output //输出分类结果
```

其中,函数 *cross_validation*(K,data)是对数据集进行 K 折交叉验证,*bestmodel* 表示第一层模型中分类效果最好的单一分类器模型,在第二层模型中做元分类器。

本文的 Stoceking 模型组合算法时间复杂度分析:假设每个单一分类模型的训练时间复杂度为 $O(M)$,由于在 stacking 模型组合过程中需使用 K 折交叉验证的方式训练每个单一分类模型,故 Stacking 模型组合算法的训练时间复杂度为 $O(\text{Models} \cdot K \cdot M)$。

4　实验结果与分析

4.1　数据集描述

选取校园网访问行为日志作为实验数据,共有 9 963 个学生 55 个属性,约 1 155.6 万条日志数据。通过日志数据的筛选、离散化得到 1 660 个学生样本及 91 个标签。其中性别、年级和年龄的人数分布如图 5 所示。

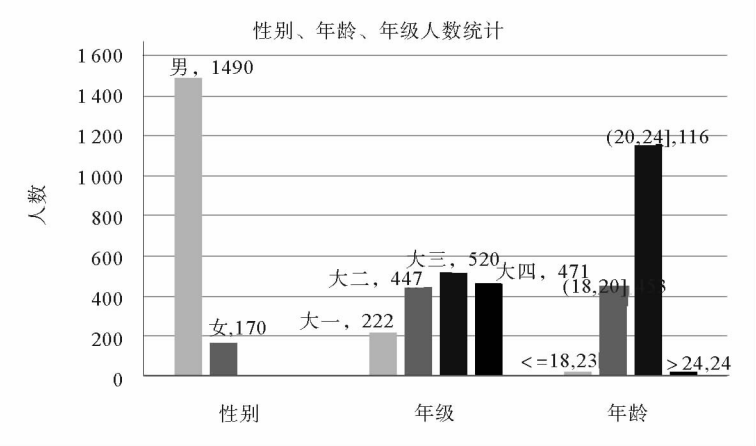

图 5　性别、年级、年龄的人数分布图

Fig. 5　Population distribution of sex, grade and age

通过图 5 可以看出性别标签中男女比例约为 9:1,说明了男生上网人数比较多。四个年级中大一上网学生人数比较少,其余三个年级上网比例较为均匀。年龄标签属性中在(20,24]年龄范围中的学生上网人数比较多。由此发现日志行为数据的属性分布不均衡,故进行二元特征提取和关联特征提取是有意义的,不仅可以扩展原始数据,而且可以避免过拟合现象。

3.2 实验结果与分析

基于用户画像模型,为了显示不同特征集合的有效性,本文采用的评价指标主要有精确率(Precision)、召回率(Recall)和 F-Measure 值(F-Measure 为 Precision 和 Recall 的加权调和均值)。

由表 1 混淆矩阵,得到精确率、召回率和 F-Measure 值的定义如下:

$$Precision=\frac{TP}{(TP+FP)}, \tag{7}$$

$$Recalll=\frac{TP}{TP+FN}, \tag{8}$$

$$F\text{-}Measure=\frac{(\alpha^2+1)Precision\times Recal}{\alpha^2 Precision+Recall}。 \tag{9}$$

表 1 混淆矩阵

Tab. 1 Confusion matrix

真实类别	预测类别	
	正例	负例
正例	TP 正确分类的正例	FN 标记为负例的正例
负例	FP 标记为正例的负例	TN 正确分类的负例

当参数 $\alpha=1$ 时,是常见的 F_1,即:

$$F_1=2\frac{Precision\times Recall}{Precision+Recall}。 \tag{10}$$

由于 F_1 集成了 *Precision* 和 *Recall* 的值,故当 F_1 较高时能更好地说明实验方法的有效性。

为验证本文提出方法的有效性,在性别、年级和年龄三个维度上进行实验。

1) 性别标签自动识别结果与分析

本组实验基于校园网行为日志数据,对性别标签进行用户画像研究。表 2 列出了基本特征下各单一分类器模型的最优实验结果。其中,RF 和 SVM 对性别分类的实验结果较好,调和平均值 F_1 均达到 0.65。在组合标签下单一分类器模型和 Stacking 组合模型最优实验结果如表 3 所示。通过表 2 与表 3 对比发现每一个单一分类器模型在组合标签上的 F_1 值均比基本标签上有较大提高,证明了特征标签可以提高对性别的分类结果。由于 RF 的结果最优,所以采用 RF 为 Stacking 组合模型第二层架构的元模型,表 3 中组合模型的 F_1 值达到了实验的最优值 0.696 9,可见,本文的用户画像方法可以提高性别预测的结果。

表 2 基本特征下单一分类器对性别的分类实验最优结果

Tab. 2 The optimal results of a single classifier for gender classification under basicfeatures

模型	*Precision*	*Recall*	F_1
SVM	0.602 9	0.758 1	0.671 7
DT	0.756 9	0.554 5	0.640 1
RF	0.598 4	0.725 4	0.655 8
LR	0.584 6	0.694 6	0.634 9
Bayes	0.619 6	0.585 2	0.601 9

表 3 特征组合下各分类器模型对性别分类的最优结果

Tab. 3 The optimal results of the classifier model for gender classification under the feature combination

模型	*Precision*	*Recall*	F_1
SVM	0.613 1	0.771 9	0.683 4
DT	0.611 7	0.761 3	0.678 3
RF	0.618 8	0.741 5	0.674 6
LR	0.608 6	0.754 4	0.673 7
Bayes	0.615 6	0.761 5	0.680 8
Stacking	0.630 7	0.718 7	0.696 9

2) 年级标签自动识别结果与分析

本组实验基于校园网行为日志数据对年级标签进行用户画像研究。表 4 和表 5 分别给出了基本方法和用户画像方法下,对学生年级分类的模型最优结果。由表 4 可以看出,在基本特征下 RF 对年级的分类效果最好,其次是 DT、LR、SVM,最差的是 Bayes。由表 5 看出,在组合标签下,DT 对年级的分类效果较好 F_1 达到0.396 4,LR 的分类效果最差。综合表 4 和表 5 可以得出:LR 对年级的分类研究结果一般;相比基本标签,组合标签下模型的训练结果的 F_1 值均有所提高,证明了组合标签有助于提高对年级的分类结果;由于单一分类器模型中 DT 的训练结果最好,故将 DT 用于组合模型中的第二层模型训练中,得到组合模型下对年级分类的 F_1 值为 0.412 1,相比单一分类器模型实验下的

最好调和均值 F_1 提升了 12.28%,证明了 Stacking 组合模型对提高年级的准确性有显著的效果,在实际应用场景中可以更准确地捕获用户属性,更好地为校园网管理工作提供精准服务。故组合多种单一分类器的 Stacking 模型是有价值的。

表 4　基本特征下单一分类器对年级的分类实验最优结果

Tab. 4　The optimal results of a single classifier for grade classification under basic features

模型	*Precision*	*Recall*	F_1
SVM	0.342 8	0.377 3	0.359 2
DT	0.361 9	0.355 9	0.358 9
RF	0.350 7	0.384 9	0.367 0
LR	0.338 4	0.359 5	0.348 6
Bayes	0.335 8	0.356 2	0.345 7

表 5　特征组合下各分类器模型对年级分类的最优结果

Tab. 5　The optimal results of classifier model for grade classification under feature combination

模型	*Precision*	*Recall*	F_1
SVM	0.373 0	0.368 6	0.370 8
DT	0.418 5	0.376 5	0.396 4
RF	0.412 0	0.371 7	0.390 8
LR	0.355 2	0.353 8	0.354 5
Bayes	0.409 3	0.368 5	0.387 8
Stacking	0.431 7	0.395 0	0.412 1

通过对网络日志样本数据集中的学生用户性别和年级的分类实验结果分析可以得出,本文提出的方法比传统分类方法可以获得更好的属性预测结果。

3) 年龄标签自动识别结果与分析

该组实验基于校园网行为日志数据对年龄标签进行用户画像研究。表 6 和表 7 分类给出了传统的单一分类器模型和本文基于 Stacking 组合模型的用户画像方法下的实验最优结果。通过对比可以发现基于组合标签下的 Stacking 组合模型,对年龄的识别结果得到了提高,证明了组合标签以及 Stacking 组合模型对提高年龄分类结果是有效的。

表 6　基本特征下单一分类器对年龄的分类实验最优结果

Tab. 6　The Optimal results of a single classifier for age classification under basic features

模型	*Precision*	*Recall*	F_1
SVM	0.259 7	0.247 5	0.243 4
DT	0.250 8	0.252 0	0.231 4
RF	0.292 7	0.320 7	0.306 1
LR	0.303 3	0.340 5	0.320 8
Bayes	0.296 3	0.308 2	0.302 1

表 7　特征组合下各分类器模型对年龄分类的最优结果

Tab. 7　The optimal results of the classifier model for age classification under the feature combination

模型	Precision	Recall	F_1
SVM	0.305 9	0.337 0	0.320 7
DT	0.302 2	0.383 8	0.338 1
RF	0.313 2	0.355 2	0.332 9
LR	0.453 2	0.335 5	0.385 6
Bayes	0.374 8	0.311 6	0.340 4
Stacking	0.487 4	0.377 8	0.425 7

4　结论

相对于传统的用户画像方法,本文提出的基于上网行为日志的用户画像方法,侧重于对用户标签进行组合,并利用 Stacking 组合模型来避免单一分类器模型的不足。通过在校园网行为日志数据上的实验分析,证明了本文所提出的基于上网行为日志的用户画像方法显著提升了对性别、年级、年龄属性的预测效果。下一步的工作中,将尝试在更大规模的数据集上,组合更多的单一分类器模型进行实验。

参考文献:

[1]费鹏. 用户画像构建技术研究[D]. 大连:大连理工大学,2017.

[2]FAWCETT T, PROVOST F. Combining data mining and machine learning for effective user profiling[J]. KDD, 1996:8-13.

[3]ADOMAVICIUS G, TUZHILIN A. User profiling in personalization applications through rule discovery and validation [C]//ACM SIGKDD International Conference on Knowledge Discovery and Data Minging DBLP,1999:377-381 .

[4]NAS RAOUI O, SOLIMAN M, SAKA E, et al. A web usage mining framework for mining evolving user profiles in dynamic web sites[J]. IEEE Transactions on Knowledge & Data Engineering,2008,26(2): 202-215.

[5]陈志明,胡震云. UGC 网站用户画像研究[J].计算机系统应用,2017,26(1):24-30.
CHEN Zhiming,HU Zhenyun. User portrait study on UGC website[J]. Computer Systems&Applications,2017,26(1):24-30.

[6]BURGER J D, HENDERSON J, KIM G, et al. Discriminating gender on Twitter[C]//Conference on Empirical Methods in N-atural Language Processing. Association for Computational Linguistics, 2011:1301-1309.

[7]FAN R E, CHANG K W, HSIEH C J, et al. Liblinear: A library for large linear classification[J]. Journal of Machine Learning Research, 2008, 9(9):1871-1874.

[8]MCCALLUM A, NIGAM K. A comparison of event models for Naive Bayes text classification[J]. In AAAI-98 Workshop on Learning for Text Categorization, 1998,62(2):41-48.

[9]CIOT M, SONDEREGGER M, RUTHS D. Gender inference of twitter users in non-English contexts[C]//EMNLP. 2013: 1136-11-45.

[10]IGLESIAS J A, ANGELOV P, LEDEZMA A, et al. Creating evolving user behavior profiles automatically[J]. IEEE Transacti-ons on Knowledge & Data Engineering, 2012, 24(5):854-867.

[11]郭光明. 基于社交大数据的用户信用画像方法研究[D].合肥:中国科学技术大学,2017.

[12]姚旭,王晓丹,张玉玺,等.特征选择方法综述[J].控制与决策,2012,27(2):161-166.
YAO Xu,WANG Xiaodan,ZHANG Yuxi,et al. Summary of feature selection algorithms[J]. Control and Decision,2012,27(2):161-166.

[13]黄源,李茂,吕建成.一种基于开方检验的特征选择方法[J].计算机科学,2015,42(5):54-56.
HUANG Yuan,LI Mao,LÜ Jiancheng. New feature selection method based on CHI[J]. Computer Science,2015,42(5):54-56.

[14]张洪强,刘光远,赖祥伟.随机森林算法在肌电的重要特征选择中的应用[J].计算机科学,2013, 40(1):200-202.
ZHANG Hongqiang,LIU Guangyuan,LAI Xiangwei. Application of random forest algorithm in important feature selection from EMG signal[J]. Computer Science,2013,40(1):200-202.

[15]施万锋,胡学钢,俞奎.一种面向高维数据的均分式 Lasso 特征选择方法[J].计算机工程与应用,2011,48(1):157-161.
SHI Wanfeng,HU Xuegang,YU Kui. k-part Lasso based on feature selection algorithm for high-dimensional data[J]. Computer Engineering and Applications,2011,48(1):157-161.

(责任编辑:傅　游)