点评用户性别预测实验报告

1753837

陈柄畅

提取特征

用户头像

使用Clarifai的API对所有用户头像进行语义分析，提取20个标签及其可能性。

```
0,21719636,{'可爱': 0.9954473972320557,
'狗': 0.992517352104187,'小巧的': 0.9882333278656006,
'动物': 0.97481369972229,'犬科': 0.9551901817321777,
'哺乳动物': 0.9496899843215942,'小狗': 0.9489724636077881,
'坐': 0.9446489810943604,'滑稽': 0.941346287727356,
'肖像': 0.9284853339195251,'一': 0.9140505790710449,
'宠物': 0.9138497710227966,'乐趣': 0.9100792407989502,
'寻找': 0.9046940803527832,'眼': 0.8945450782775879,
'沒有人': 0.8931015729904175,'毛皮': 0.8873289823532104,
'年轻': 0.8711099624633789,'幽默': 0.8442040681838989,
'工作室': 0.8420847058296204}
```

用户名

参考论文 中文微博用户性别分类方法研究，使用Jieba对用户名进行分词，提取首字、尾字并根据HIT中文停用词和Github上的英文停用词过滤掉停用词和出现次数为一次的词。

```
5,15582434,"['晚_f', '荷_l', '晚', '荷']"
6,1026259143,"['麦_f', '__l', '麦', '兜', '不吃', '粗', '_']"
7,39110333,"['明_f', '号_l', '明天', '32', '号']"
8,131367256,"['陈_f', 'C_l', '陈', '珍妮', 'Jenny.C']"
9,7152956,"['b_f', '7_l', 'blackberry77']"
```

用户评论

使用gensim对每条评论进行Doc2Vec向量化，最后对每个用户的所有评论向量进行平均作为用户的特征向量。

```
1,21719636,"[ 5.70256665e-02 -2.70835325e-01 -3.29258536e-01 -2.47345232e-01
  1.22436173e-01 -3.03500598e-01  1.52676687e-01  9.87093743e-02
 -2.37379018e-02 -9.83558347e-02 -6.48951841e-01 -3.40015550e-01
  4.00736133e-02 -1.40577491e-01 -3.90931786e-02  1.80064548e-01
 -4.89567774e-01  8.17305798e-02 -8.25695098e-02  9.53902793e-02
 -2.57575148e-01  4.61537841e-01 -2.02138397e-01 -2.98381398e-01
  8.52721825e-02  1.64483529e-01 -1.68684073e-01  3.22851369e-01
  4.35252245e-02 -2.03261443e-01  7.23548916e-01  2.03579291e-01
  1.32416540e-01  1.39608649e-01  2.02670484e-01 -2.60425602e-01
 -2.07082453e-01 -2.07509994e-01 -7.73019792e-01 -1.96182879e-02
 -3.42726428e-02  1.86046738e-02  9.77607119e-02 -1.46981390e-01
 -3.85673396e-02 -3.16003497e-01 -8.08042725e-02 -8.01830612e-02
  3.28016416e-05  2.06488212e-01  1.62513190e-01  2.74405069e-01
  8.22204003e-02  2.38853496e-01 -1.48990202e-02 -4.82907030e-03
 -6.24884875e-01  2.88120528e-01 -2.83794640e-01  2.35593099e-01
  2.58314739e-01 -2.96983622e-01  2.33479642e-02  4.56174395e-01
  1.20707182e-01 -1.97471379e-01  5.90253766e-03 -2.65562230e-01
 -3.47100672e-02  2.30598583e-01  1.51274162e-01  1.06508115e-01
 -1.18816902e-01  2.71618441e-01 -6.51841323e-01 -3.79984084e-01
 -4.97678543e-01 -5.14618068e-03  2.64919258e-01  1.20109942e-01
  3.81864658e-01 -1.23843767e-01 -9.40665284e-01 -2.51718718e-01
 -2.62600899e-01  5.93655896e-01 -6.46966181e-02 -1.37339051e-01
  4.58858945e-01  1.58114528e-01 -2.47366831e-01  1.58034485e-01
 -1.52278344e-01  1.85399736e-01 -5.25616629e-02  2.80527996e-01
 -1.98700496e-01 -1.91485554e-01 -2.62768990e-01  1.48365708e-01]"
```

使用SnowNLP对每条评论进行情感分析，最后平均作为用户的特征向量。

```
0,21719636,0.6139398611162253
1,2574437,0.7137682998241126
2,1964659,0.6837056250882225
3,2774623,0.452256761355288
4,3861727,0.6262338703761681
5,15582434,0.708180812400469
```

提供数据的再处理

- 将用户向量展开
- 对`figuration`转换为平均每篇评论中出现的次数
- 将`word_count`转换为平均每篇评论字数及用户发表评论数

上述数据以及训练集、验证集、测试集均data文件夹中。

数据预处理

在Logistic Regression, SVM, Random Forest分类之前对数据进行了归一化处理。

在Neural Network (多层感知器) 分类之前，由于`solver`使用`adam`, 所以对数据进行了标准化处理。

特征选取

本次实验分为提取特征、提供的特征和所有特征（由于提取特征和提供的特征中对于review向量化重复，故只选取Doc2Vec）三种特征选取方式进行实验。

另外，在所有特征的实验中，分别删除评论情感，表情符号统计两个特征，分类准确度都有所降低。

参数设定

SVM

训练样本的特征数量巨大，不需要通过RBF等非线性核函数将其映射到更高的维度空间上，利用非线性核函数也并不能提高分类器的性能。利用linear核函数就可以获得足够好的结果。

Random Forest

尝试了在验证集调整一系列参数，但效果都不如默认参数好。。。

Neural Network

- 自己提取的特征

```
MLPClassifier(hidden_layer_sizes=
[100,100,10],max_iter=1000,random_state=33,alpha=5)
```

三层隐藏层，每层分别100、100、10个神经元，正则化系数为5

- 提供的特征

```
MLPClassifier(hidden_layer_sizes=[20,20,10],max_iter=1000,random_state=33)
```

三层隐藏层，每层分别20、20、10个神经元，正则化系数为0.0001

- 所有的特征

```
MLPClassifier(hidden_layer_sizes=
[100,10,10],max_iter=1000,random_state=33,alpha=1)
```

三层隐藏层，每层分别100、10、10个神经元，正则化系数为1

由于设备限制，神经网络只选择了三层进行了调参。

实验结果

参考论文Cross-domain gender detection in Twitter, 使用stacking作为集成学习模型，meta classifier使用`Logistic Regression`。

	My Feature	Given Feature	All
Logistic Regression	Scores on test dataset Accuracy: 0.8219248584662893 Precision: 0.8605072463768116 Recall: 0.9253246753246753 F1: 0.8917396745932414 Scores on train dataset Accuracy: 0.9053587415270495 Precision: 0.9071330380326867 Recall: 0.9782572887497941 F1: 0.9413536218101125	Scores on test dataset Accuracy: 0.8279166666666666 Precision: 0.8451612903225807 Recall: 0.9440133037694013 F1: 0.8918565069389893 Scores on train dataset Accuracy: 0.8319616626731952 Precision: 0.8451888094341989 Recall: 0.9467397414277684 F1: 0.8930867634387221	Scores on test dataset Accuracy: 0.8172928461142563 Precision: 0.8580060422960725 Recall: 0.922077922077922 F1: 0.888888888888889 Scores on train dataset Accuracy: 0.9045913799718635 Precision: 0.906054598139393 Recall: 0.9785867237687366 F1: 0.9409249287298067
Logistic Regression + Stacking	Scores on test dataset Accuracy: 0.8219248584662893 Precision: 0.8605072463768116 Recall: 0.9253246753246753 F1: 0.8917396745932414 Scores on train dataset Accuracy: 0.9053587415270495 Precision: 0.9071330380326867 Recall: 0.9782572887497941 F1: 0.9413536218101125	Scores on test dataset Accuracy: 0.8279166666666666 Precision: 0.8451612903225807 Recall: 0.9440133037694013 F1: 0.8918565069389893 Scores on train dataset Accuracy: 0.8319616626731952 Precision: 0.8451888094341989 Recall: 0.9467397414277684 F1: 0.8930867634387221	Scores on test dataset Accuracy: 0.8172928461142563 Precision: 0.8580060422960725 Recall: 0.922077922077922 F1: 0.888888888888889 Scores on train dataset Accuracy: 0.9045913799718635 Precision: 0.906054598139393 Recall: 0.9785867237687366 F1: 0.9409249287298067
SVM	Scores on test dataset Accuracy: 0.8095728255275347 Precision: 0.8620049504950495 Recall: 0.9045454545454545 F1: 0.8827629911280102 Scores on train dataset Accuracy: 0.928251694590101 Precision: 0.931412464766677 Recall: 0.9797397463350355 F1: 0.9549650798747692	Scores on test dataset Accuracy: 0.8320833333333333 Precision: 0.8500749625187406 Recall: 0.9429046563192904 F1: 0.8940867279894875 Scores on train dataset Accuracy: 0.8380039587457027 Precision: 0.8510289104910996 Recall: 0.9473018549747049 F1: 0.8965884152424021	Scores on test dataset Accuracy: 0.8064848172928462 Precision: 0.8619402985074627 Recall: 0.9 F1: 0.8805590851334181 Scores on train dataset Accuracy: 0.928251694590101 Precision: 0.932224662692187 Recall: 0.9787514412782079 F1: 0.9549216552832464
SVM + Stacking	Scores on test dataset Accuracy: 0.8095728255275347 Precision: 0.8620049504950495 Recall: 0.9045454545454545 F1: 0.8827629911280102 Scores on train dataset Accuracy: 0.928251694590101 Precision: 0.931412464766677 Recall: 0.9797397463350355 F1: 0.9549650798747692	Scores on test dataset Accuracy: 0.8320833333333333 Precision: 0.8500749625187406 Recall: 0.9429046563192904 F1: 0.8940867279894875 Scores on train dataset Accuracy: 0.8380039587457027 Precision: 0.8510289104910996 Recall: 0.9473018549747049 F1: 0.8965884152424021	Scores on test dataset Accuracy: 0.8064848172928462 Precision: 0.8619402985074627 Recall: 0.9 F1: 0.8805590851334181 Scores on train dataset Accuracy: 0.928251694590101 Precision: 0.932224662692187 Recall: 0.9787514412782079 F1: 0.9549216552832464
Random Forest	Scores on test dataset Accuracy: 0.7848687596500258 Precision: 0.8311688311688312 Recall: 0.9142857142857143 F1: 0.8707482993197279 Scores on train dataset Accuracy: 0.9934774267809183 Precision: 0.9924738219895288 Recall: 0.9991764124526438 F1: 0.9958138389559222	Scores on test dataset Accuracy: 0.7666666666666667 Precision: 0.8315565031982942 Recall: 0.8647450110864745 F1: 0.8478260869565217 Scores on train dataset Accuracy: 0.994270236482967 Precision: 0.9960657580441197 Recall: 0.9962057335581788 F1: 0.9961357408838614	Scores on test dataset Accuracy: 0.7925887802367473 Precision: 0.8338226658837345 Recall: 0.922077922077922 F1: 0.875732346592661 Scores on train dataset Accuracy: 0.9929658524107943 Precision: 0.9923076923076923 Recall: 0.99868225992423 F1: 0.995484771365241
Random Forest + Stacking	Scores on test dataset Accuracy: 0.8095728255275347 Precision: 0.8286516853932584 Recall: 0.9577922077922078 F1: 0.8885542168674697 Scores on train dataset Accuracy: 0.999744212814938 Precision: 0.9996706734727482 Recall: 1.0 F1: 0.9998353096179182	Scores on test dataset Accuracy: 0.7920833333333334 Precision: 0.8196962273395394 Recall: 0.9273835920177383 F1: 0.8702210663198958 Scores on train dataset Accuracy: 0.9994791124075425 Precision: 0.9994381233319286 Recall: 0.9998594716132658 F1: 0.9996487530734106	Scores on test dataset Accuracy: 0.8023674729799279 Precision: 0.8239910313901345 Recall: 0.9545454545454546 F1: 0.884476534296029 Scores on train dataset Accuracy: 0.9989768512597519 Precision: 0.9988481158466348 Recall: 0.9998352824905288 F1: 0.9993414553836023
LR + SVM +RF + Stacking	Scores on test dataset Accuracy: 0.8286155429747812 Precision: 0.8354641467481935 Recall: 0.9759740259740259 F1: 0.9002695417789757 Scores on train dataset Accuracy: 0.9457731167668499 Precision: 0.9347190146266359 Recall: 1.0 F1: 0.9662581569314023	Scores on test dataset Accuracy: 0.8308333333333333 Precision: 0.846382556987116 Recall: 0.9467849223946785 F1: 0.8937728937728938 Scores on train dataset Accuracy: 0.8410251067819564 Precision: 0.8515723270440252 Recall: 0.9513771781899943 F1: 0.8987123324040887	Scores on test dataset Accuracy: 0.8172928461142563 Precision: 0.8554289142171566 Recall: 0.925974025974026 F1: 0.8893046460866855 Scores on train dataset Accuracy: 0.9345184806241207 Precision: 0.9342290267145759 Recall: 0.9850107066381156 F1: 0.9589480436177036
Neural Network	Scores on test dataset Accuracy: 0.8244981986618631 Precision: 0.8288535381239714 Recall: 0.9811688311688311 F1: 0.8986024382991377 Scores on train dataset Accuracy: 0.819542140938739 Precision: 0.8179585152838428 Recall: 0.9873167517707132 F1: 0.8946936338532726	Scores on test dataset Accuracy: 0.8254166666666667 Precision: 0.8654353562005277 Recall: 0.9090909090909091 F1: 0.8867261422005949 Scores on train dataset Accuracy: 0.8446713199291593 Precision: 0.8749500066657779 Recall: 0.9222878021360315 F1: 0.8979954847095847	Scores on test dataset Accuracy: 0.8353062274832733 Precision: 0.8530092592592593 Recall: 0.9571428571428572 F1: 0.9020807833537333 Scores on train dataset Accuracy: 0.86507225987978 Precision: 0.8646408839779005 Recall: 0.9795750288255641 F1: 0.9185265271449533

由于采用了多种方法，实验结果文件无法按照作业要求的格式进行展示。故采用下列格式：

user_id	lr_label	lrs_label	svm_label	svms_label	rf_label	rfs_label	mix_label	nn_label	true_label
用户id	Logistic Regression 预测结果	Logistic Regression + Stacking预测结果	SVM预测结果	SVM + Stacking预测结果	Random Forest预测结果	Random Forest + Stacking预测结果	LR + SVM +RF + Stacking预测结果	Neural Network 预测结果	实际值

比较分析

- 集成学习对于多个相同的基分类器的提升效果不大
- 由于随机森林本身就是集成学习模型，stacking对其准确度有所提升，但也增加了其过拟合的程度。
- 神经网络在分类时也极易发生过拟合的现象，但可以通过简化神经网络和增大正则化系数进行减少。
- 对于本数据集，LR + SVM +RF + Stacking 和 Neural Network 的方法分类效果最好。时间效率上看，SVM + Stacking的方法效率最低。
- 同一算法对不同特征的表现有所差异，不同算法分别使用适用于不同特征，但LR + SVM +RF + Stacking 和 Neural Network在不同特征集上的效果都很不错。

改进

由于时间和设备的限制，本次实验仍有一些可以改进的地方。

- 随机森林存在着过拟合的情况，但计算设备有限，没有能够调整参数到合理水平。
- 尝试对于特征选取进行更加细致的调试
- 尝试其他集成学习模型
- 尝试不同的神经网络模型
- 尝试不同的meta classifier