计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(2):240−244 [doi: 10.15888/j.cnki.csa.006206]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于多视图 Tri-Training 的微博用户性别判断①

孙启蕴

(南京烽火软件科技有限公司, 南京 210019)
(武汉邮电科学研究院 通信与信息专业, 武汉 430073)
通讯作者: 孙启蕴, E-mail: 273048269@qq.com

摘　要: 互联网技术不断发展, 新浪微博作为公开的网络社交平台拥有庞大的活跃用户. 然而由于用户数量庞大, 且个人信息并不一定真实, 造成训练样本打标困难. 本文采用了一种多视图 tri-training 的方法, 构建三个不同的视图, 利用这些视图中少量已打标样本和未打标样本不断重复互相训练三个不同的分类器, 最后集成这三个分类器实现用户性别判断. 本文用真实用户数据进行实验, 发现和单一视图分类器相比, 使用多视图 tri-training 学习训练后的分类器准确性更好, 且需要打标的样本更少.

关键词: 性别判断; 多视图学习; tri-training 算法; 数据挖掘

引用格式: 孙启蕴.基于多视图 Tri-Training 的微博用户性别判断.计算机系统应用,2018,27(2):240–244. http://www.c-s-a.org.cn/1003-3254/6206.html

Microblog User Gender Recognition with Multi-View and Tri-Training Learning

SUN Qi-Yun

(FiberHome Telecommunication Technologies Co. Ltd., Nanjing 210019, China)
(Wuhan Research Insititute of Posts and Telecommunications, Wuhan 430073, China)

Abstract: With the high pace of internet technology, microblog, an opening free social network, has an awful lot of active users. However, the number of sina microblog users is very large and the personal information is not always true, leading to the situation that it is hard to label the user's gender. In this study, multi-view and tri-training learning method are used to solve these problems. First three different views are constructed and three different classifiers are trained with a small number of labeled samples. And then three different classifiers are trained repeatedly by unlabeled samples. Finally, we integrate three classifiers into one to judge the user gender. We use the real user data and find that the classifier using the multi-view and tri-training learning is better than the performance of the single view classifier and needs less labeled data.

Key words: gender recognition; multi-view learning; tri-training; data mining

引言

随着人们对互联网的使用量逐渐增大, 互联网进入了大数据时代. 新浪微博作为一个公开社交平台, 使人们能够有获取最新最热门的新闻、了解话题舆论、展现自我观点、寻找志趣相投的朋友等途径. 截止到 2017 年第一季度, 新浪微博约有 2.97 亿日常活跃用户, 每天新增的微博数量约 4 亿条. 而新浪微博用户性别这一基本属性在不同领域都有着重要影响, 如微博推荐系统会根据性别的不同给男性推荐车、体育相关的微博, 而给女性推荐美妆、衣服相关的微博等等. 因此对新浪微博用户的性别做判断很有意义.

目前国内外有不少研究人员对社交平台中的用户性别判断进行研究. 文献[1]对 twitter 中非英文用户性别的特征进行探索; 文献[2]利用用户间的评论信息文

① 收稿时间: 2017-05-17; 修改时间: 2017-06-16; 采用时间: 2017-06-19

本推测出交互双方的性别; 文献[3]通过一个分类器寻找两个博客之间的关系, 从关联来获取未标注样本中的分类信息; 文献[4]采用半监督学习方法, 通过博客文本和博客评论两个视图对用户性别进行分类, 取得了不错的分类性能. 文献[5]从两性表达情绪的差异出发, 利用微博发布文本内容中的情绪特征进行性别判断; 文献[6]将用户兴趣标签分成若干概念类来区分用户性别, 但这两篇文献在实践过程中都需要大量人工标记样本, 且准确度不高.

本文从新浪微博爬取真实的用户数据, 经过前期数据清洗过滤之后, 利用微博文本信息、微博用户标签、微博用户昵称三个可以刻画微博用户性别的方向作为三个不同的视图, 采用 tri-training 算法对三个不同的分类器进行互相训练学习. 实验结果表明, 在只用少量已标注训练集的情况下, 多视图 tri-training 学习能有效的提高分类器的性别分类效果.

1 Tri-training 和多视图

在传统的机器学习分类问题中, 一般分为有监督和无监督两类学习方法[7]. 随着大数据时代的发展, 我们往往获得的是大量未标记数据和少量已标记过的数据. 在训练时, 如果不考虑大量未标记的数据将会造成有用信息的丢失, 同样, 如果只用少量已标记数据训练, 很难保证训练器的准确性. 半监督学习[8]利用大量未标记数据和少量已标记数据对训练器进行训练, 省去了人工打标的时间同时提高了分类器的性能.

在主流的半监督学习算法中, 最具代表性的就是协同训练 (co-training), 它提出[9]如果数据集中有两个充分冗余的视图, 那么分别用两个视图上已记数据各自训练处一个分类器然后在协同训练时, 每个分类器从未标记数据中选择置信度较高的数据进行标记, 这样另一分类器就可以根据这些新标记的数据重新进行训练. 这样两个分类器能通过互相训练未知信息, 使得自身准确性更高.

本文采用半监督学习中的 tri-training 算法, 通过三个不同的分类器之间相互学习训练来处理未知类别分类问题. 与 co-training 不同, tri-training 算法采用了非显示投票来处理置信度, 在最初的分类器分类准确还很低的时候, 辅助分类器对未打标数据的判断可能会同时判断成其他的类别, 从而引入噪音[10]. 噪音学习理论[11]中提到, 如果辅助分类器能正确的判断大部分未标记训练数据, 那么噪声所带来的错误率会被抵消. 因此在不断重复训练分类器时, 只要保证下一次的分类误差率小于本次的分类误差率就认为训练过程正常. 直到下一次的分类误差率大于本次的分类误差率, 那么分类器训练结束.

$$0 < \frac{e_{l+1}}{e_l} < \frac{|L_l|}{|L_{l+1}|} < 1 \tag{1}$$

l 为训练循环了第 l 次, e_l 为第 l 次训练过程中的误差, L_l 为第 l 次训练过程中已打标样本和另两个分类器对未打标样本分类相同的集合.

当满足表达式 (1) 的时候, 就能保证下一次的分类误差率小于本次的分类误差率, 未标记数据集可以作为训练样本对分类器进行训练, 使得大量新样例加入到初始训练集对分类器进行重复的训练, 从而使引入噪声所带来的负面影响被大量的未标记数据所带来的好处抵消[12].

尽管半监督学习已经研究了十几年, 但是仍有其局限性, 他们研究的数据只有一个特征集, 忽略了大数据的异构性, 会造成信息的丢失[13]. 现实情况中对象存在多个视图, 刻画一个事物能通过不能的角度或者通过不同的工具[14]. 通常可以用 (x_i, y_i) 来表示用单视图描述的对象, 其中 x_i 是一个对象, y_i 是确定类别的标签. 而我们用 $([x_{i1}, x_{i2}, x_{i3}], y_i)$ 来表示一个多视图的对象, 其中$[x_{i1}, x_{i2}, x_{i3}]$是用一些不同视图来刻画同一个对象 (比如多媒体数据, x_{i1} 为文本视图, x_{i2} 为图像视图, x_{i3} 为视频视图). 虽然在进行协同训练的时候并不一定需要多视图, 但是多视图往往有锦上添花的能力. 文献[15]指出, 在冗余的多视图上, 由于视图之间有着有用信息, 即使只用一个已标记数据作为起始训练样本, 半监督学习也能顺利的进行下去.

2 微博用户性别判断方法

本文采用基于多视图 tri-training 学习的途径来判断性别.

我们先对三个视图 (微博文本信息视图、微博用户标签信息视图、微博用户昵称视图) 建立维度特征, 然后对这三个不同的视图分别训练三个不同的分类器并使它们互相学习训练未标记的样本数据, 最后将已经训练好的三个分类器进行集成, 来对测试样本进行分类. 微博用户性别判断流程图如图 1 所示.

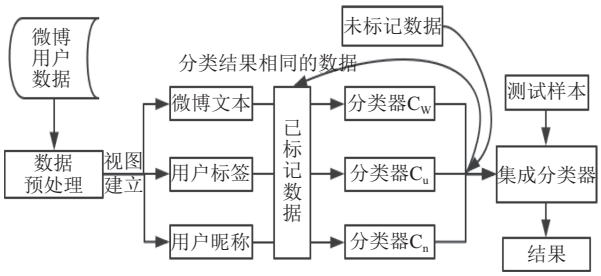

图 1 微博用户性别判断流程图

2.1 多视图分析与建立

2.1.1 微博文本信息视图

微博文本信息在一定程度上能反映出用户的性别, 男性用户在表达感情上更喜欢用表达愤怒厌恶的情绪相关的词汇. 而女性微博的发言更可能会出现“嗨皮啊！[亲亲][亲亲]希望男神的新专辑大麦！小女子支持到底！”包含“男神”、“小女子”以及连续重复表情符“[亲亲]”等词汇.

本文在处理微博文本信息上先进行分词、去停用词的操作, 然后采用向量空间模型 (VSM) 把文本转换成空间向量. 未做任何处理的空间向量由文本切分的所有词组成, 如果不降维会产生维度灾难. 因此需要对微博文本信息进行降维处理, 本文采用信息增益 (IG) 来进行特征选择.

IG 的重要衡量条件就是判断该特征能带来多大的信息量, 信息越多则表明该特征越重要. 如一个特征 f, 有该特征的信息量与没有该信息的信息量差值即为 f 的信息增益. 另外, 在降维处理的幅度上, 若减少的维度过多, 会影响分类器的准确性, 若特征数仍过多, 会存在很多噪音. 本文的特征选择 IG 最高的前 5000 个.

2.1.2 微博用户标签视图

微博用户标签是微博用户根据喜好或者自身属性而打上的标签, 这些标签能反映出用户在当前阶段的兴趣、关注点和自身情况. 据统计, 约有 53% 的用户会添加自己的标签.

从表 1 微博用户标签信息举例可以看出: 女性用户的标签信息中往往会带有透露自己性别的字眼, 如“妞儿”、“女金牛”等, 且往往标签不止一个兴趣词来描述自己, 而是会增加一些形容词如“能吃的”、“不脑残的”, 或者表示程度的副词“很”等等. 而男性用户的标签大多仅为简短的兴趣词汇, 并未出现同表达程度和感情的形容词或者副词. 因此在特征选择时, 加入程度词频率及标签平均长度这两个维度.

表 1 微博用户标签信息举例

性别	用户标签
女	傻丫头 大懒虫 笑点低 爱喝的妞儿 脑残儿童欢乐多
女	不爱动的胖子 喜欢安静 爱冷幽默 女金牛 爱粤语歌爱港剧 不脑残的90后 爱旅游 很能吃的吃货 很会睡 居里夫人一枚
男	金融IT运营 网站运营 单身 上海
男	白羊男 公益 宅 音乐 80后 旅游 听歌 电影

2.1.3 微博用户昵称视图

微博用户昵称并非实名制, 用户可以按照自己的喜好和兴趣或者情绪来创建昵称. 虽然没有限制条件, 但是用户在取名的时候仍会受到性别的影响. 如“叶仁琛”、“老男孩不加 V”、“HelloWorld 天真浪子”等男性化的词汇更可能为男性用户的昵称, 而女性用户的昵称更可能出现“沐雪莹莹”、“高姿态的妞儿”、“捣蛋_女孩”等女性化词汇.

与微博文本信息不同, 由于微博用户昵称字数较短, 使用分词可能会造成昵称无法被正确切分, 因此在对用户昵称的提取上采用 n-Gram 来提取特征来避免切词障碍. 我们选择 n-Gram 中 n=1 和 n=2, 即 unigram 和 bigram 两种特征提取方式. 其中 unigram 为一元字特征, bigram 为二元字特征. 表 2 列举了微博用户昵称“高姿态的妞儿”和“叶仁琛”分别用 unigram、bigram、unigram+bigram 和结巴中文分词进行特征提取的结果.

表 2 微博用户昵称文本特征举例

	“高姿态的妞儿”	“叶仁琛”
Unigram	“高”、“姿”、“态”、“的”、“妞”、“儿”	“叶”、“仁”、“琛”
Bigram	“高姿”、“姿态”、“态的”、“的妞”、“妞儿”	“叶仁”、“仁琛”
Unigram+Bigram	“高”、“姿”、“态”、“的”、“妞”、“儿”、“高姿”、“姿态”、“态的”、“的妞”、“妞儿”	“叶”、“仁”、“琛”、“叶仁”、“仁琛”
结巴分词	“高姿态”、“的”、“妞儿”	“叶仁琛”

2.2 改进的 tri-training 训练分类器

三个视图分别为微博文本信息、微博用户标签信息、微博用户昵称, 经过 tri-training 算法后生成三个不同的分类器, 分别为微博文本分类器 Cw、用户标签分类器 Cu、用户昵称分类器 Cn. 由于传统 tri-training 训练的基分类器均为同一类型的监督学习分类, 泛化效果不理想[16], 而且多视图的内容各不相同, 如果使用同一种类型的分类器, 可能对于某几个视图该种分类

器相比于其他类型分类器的分类性能弱. 因此本文在传统算法的基础上, 针对每个视图的特征特点来选取不同的监督学习分类器. 由于 SVM 分类器能很好的解决在小样本情况下高维模型的问题, 本文在用户标签视图分类器 Cu 选择 SVM 分类器; 而最大熵分类器融合信息的能力较好, 可以解决较复杂的问题, 因此在微博文本视图分类器 Cw 和在用户昵称视图分类器 Cn 选择最大熵分类器.

算法流程如下:

输入:

原始已标记数据集 L={微博文本 Lw、用户标签 Lu、用户昵称 Ln}

原始未标记数据集 U={微博文本 Uw、用户标签 Uu、用户昵称 Un}

输出:

微博文本分类器 Cw、用户标签分类器 Cu、用户昵称分类器 Cn

步骤:

1) 使用 Lw、Lu、Ln 分别对初始分类器进行训练, 得到 Cw、Cu、Cn;

2) 对每个分类器分别进行以下步骤直到满足指定条件时停止 (下面以 Cw 为例);

3) 将 Ui 中的用户标签 Uui、Uni 分别放入分类器 Cu、Cn 进行分类;

4) 将 Cu、Cn 分类结果相同的 Ui 中的 Uwi 和 Lw 组合成新的训练样本 Lw';

5) 使用 Lw'重新训练分类器 Cw;

6) 将 Uwi 重新放回 Uw 中进行下一轮的分类;

7) 当新分类器的迭代指定次数时或者原始未标记数据集 U 为空时终止.

2.3 分类器集成

传统分类器的集成往往通过简单投票法[17], 比如三个中如果有 2 个分类器的结果相同那么就判定为该类别. 但是这种方法在融合的时候没有考虑到三个分类器自身分类强弱特性, 当其中一个较强分类器判断正确, 另两个分类器判断错误时, 会出现较大偏差导致最后的结果分类错误. 因此本文在使用 tri-training 训练结束生成三个视图的分类器后, 以准确率作为权重对这三个分类器进行集成, 准确度越高的分类器的权重就越大. 这样能在分类器的分类性能存在差异的时, 使判断的结果更加准确.

3 实验结果

本文实验数据均来自真实新浪微博用户数据, 使用 python 脚本爬虫爬取 15 000 名用户的微博文本、用户标签和用户昵称. 并对内容做出限制, 筛选出微博文本条数大于 30 条, 用户标签大于 4 个的非企业认证 (蓝 V) 用户共 6841 名.

由于 6841 名微博用户是随机爬取, 因此在实验前先人工对这些微博用户进行打标, 根据其微博文本、标签、昵称、相册和评论来判定其性别, 最后选出男女用户各 2500 名, 共计 5000 名. 本文选取 20% 的数据 (1000 名用户) 作为测试样本集, 80% 的数据 (4000 名用户) 作为训练样本集. 其中选取训练样本的 30% 作为已打标数据, 剩下 70% 作为未打标数据.

本文比较单一视图下使用有限标记样本进行监督学习的分类器和使用多视图 tri-training 学习后三个分类器的检测准确度的差异, 并比较了利用本文算法集成后的分类器准确度, 实验结果如图 2 所示. 从图 2 可以看出基于多视图 tri-training 学习后的分类器判断效果更好, 并且按照准确度权重进行集成后的分类器准确度提高了 1%.

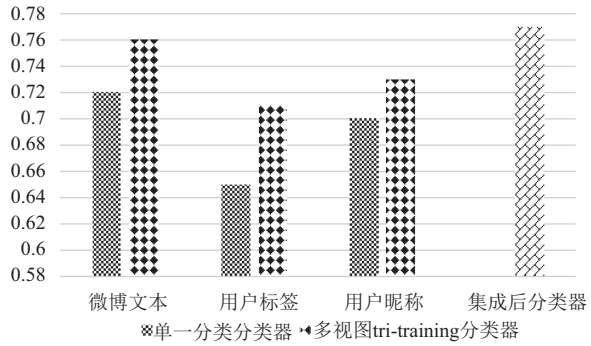

图 2 单一分类器和多视图 tri-training 准确性比较

另外, 在三个视图的分类器选择上, 和传统的 tri-training 采用同一种分类器不同, 本文在比较多种分类器组合后选择使用一个 SVM 分类器和两个最大熵分类器. 多种分类器组合情况和比较的结果如表 3 和表 4 所示.

表 3 多种分类器组合情况

	微博文本视图	用户标签视图	用户昵称视图
GROUP1	SVM	SVM	SVM
GROUP2	最大熵	最大熵	最大熵
GROUP3	最大熵	SVM	最大熵

表 4 多种分类器组合准确性比较

	GROUP1	GROUP2	GROUP3
准确性	0.743	0.765	0.771

从对比可以看出, 在对微博用户性别进行判断时, 多视图 tri-training 学习得到的分类器性能比单视图分类器效果更好. 而且在分类器的选择上, 三个视图合各自特征选择合适的分类器组合比三个视图使用同一分类器准确度更高.

4 结束语

本文结合多视角学习和半监督学习的方法, 在大量新浪微博用户性别数据打标困难的情况下, 通过少量人工打标样本和大量未标记样本, 利用微博文本、用户标签、用户昵称三个视图对三个分类器相互学习训练. 通过真实用户数据实验后, 发现多视图学习后的分类器在对微博用户性别进行分类的准确性上比单一视图分类器效果更好. 但本文在实验过程中只从三个视图出发对用户性别做判断, 而微博中的话题、评论、关注人等都能在一定程度上体现出用户性别, 今后可以尝试从更多角度判断用户性别.

参考文献

1 Ciot M, Sonderegger M, Ruths D. Gender inference of twitter users in Non-English contexts. Stroudsburg. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, WA, USA. 2013. 1136–1145.

2 Li SS, Wang JJ, Zhou GD, *et al*. Interactive gender inference with integer linear programming. Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina. 2015. 2341–2347.

3 Ikeda D, Takamura H, Okumura M. Semi-supervised learning for blog classification. Proceedings of the 23rd National Conference on Artificial Intelligence. Chicago, IL, USA. 2008. 1156–1161.

4 Wang JJ, Xue YX, Li SS, *et al*. Leveraging interactive knowledge and unlabeled data in gender classification with co-training. International Conference on Database Systems for Advanced Applications. Hanoi, Vietnam. 2015. 246–251.

5 刘宝芹, 牛耘. 基于情绪特征的中文微博用户性别识别. 计算机工程与科学, 2016, 38(9): 1917–1923.

6 钱铁云, 尤珍妮, 陈丽, 等. 基于兴趣标签的缄默用户性别预测研究. 华中科技大学学报 (自然科学版), 2015, 43(12): 101–105.

7 蓝超, 饶泓, 浣军. 半监督多视图学习在大数据分析中的应用探讨. 中兴通讯技术, 2015, 21(5): 32–34.

8 Yin CY, Xiang J, Zhang H, *et al*. A new SVM method for short text classification based on semi-supervised learning. Proceedings of International Conference on Advanced Information Technology and Sensor Application. Harbin, China. 2015. 100–103.

9 郭翔宇, 王魏. 一种改进的协同训练算法: Compatible Co-training. 南京大学学报 (自然科学), 2016, 52(4): 662–671.

10 兰霞. 半监督协同训练算法的研究[硕士学位论文]. 成都: 四川师范大学, 2011.

11 闫耀辉, 臧洌, 黄同心. 基于协同训练的 Co-Forest 算法在入侵检测中的应用. 2010 通信理论与技术新发展——第十五届全国青年通信学术会议论文集 (下册). 昆明, 中国. 2010. 305–309.

12 Sun SL. A survey of multi-view machine learning. Neural Computing and Applications, 2013, 23(7-8): 2031 –2038. [doi: 10.1007/s00521-013-1362-6]

13 Xu C, Tao DC, Xu C. A survey on multi-view learning. arXiv:1304.5634, 2013: 1–49.

14 于重重, 刘宇, 谭励, 等. 组合标记的多视图半监督协同分类算法. 计算机应用, 2013, 33(11): 3090–3093.

15 Qian TY, Liu B, Chen L, *et al*. Tri-Training for authorship attribution with limited training data: A comprehensive study. Neurocomputing, 2016, 171: 798–806. [doi: 10.1016/j.neucom.2015.07.064]

16 Chou CL, Chang CH, Huang YY. Boosted web named entity recognition via tri-training. ACM Transactions on Asian and Low-Resource Language Information Processing, 2016, 16(2): 10.

17 张荣荣. 图像分类中融合 Bagging 的 Tri-Training 算法研究[硕士学位论文]. 重庆: 西南大学, 2016.