DOI:10.7544/issn1000-1239.2015.20131473
52(3):639-648, 2015

基于 PU 学习算法的虚假评论识别研究

任亚峰　姬东鸿　张红斌　尹　兰
(武汉大学计算机学院　武汉　430072)
(renyafeng@whu.edu.cn)

Deceptive Reviews Detection Based on Positive and Unlabeled Learning

Ren Yafeng, Ji Donghong, Zhang Hongbin, and Yin Lan
(*School of Computer*, *Wuhan University*, *Wuhan* 430072)

Abstract　Identifying deceptive reviews has important theoretical meaning and practical value. While previous works focus on some heuristic rules or traditional supervised methods. Recent research has shown that humans cannot directly identify deceptive reviews by their prior knowledge. Human-annotated dataset must contain some mislabeled examples. Due to the difficulty of human labeling needed for supervised learning, the problem remains to be highly challenging. There are some ambiguous reviews (we call them spy examples), which are easily mislabeled. The key of identifying deceptive review is how to deal with these spy reviews. Based on some truthful reviews and a large amount of unlabeled reviews, a novel approach, called mixing population and individual nature PU learning, is proposed. Firstly, some reliable negative examples are identified from the unlabeled dataset. Secondly, some representative positive examples and negative examples are generated by integrating latent dirichlet allocation and *K*-means. Thirdly, all spy examples are clustered into many groups based on dirichlet process mixture model, and two schemes (population nature and individual nature) are mixed to determine the category label of spy examples. Finally, multiple kernel learning is presented to build the final classifier. Experimental results demonstrate that our proposed methods can effectively identify deceptive reviews, and outperform the current baselines.

Key words　deceptive reviews; supervised learning; positive and unlabeled (PU) learning; Dirichlet process mixture model (DPMM); multiple kernel learning (MKL)

摘　要　识别虚假评论有着重要的理论意义与现实价值.先前工作集中于启发式策略和传统的全监督学习算法.最近研究表明:人类无法通过先验知识有效识别虚假评论,手工标注的数据集必定存在一定数量的误例,因此简单使用传统的全监督学习算法识别虚假评论并不合理.容易被错误标注的样例称为间谍样例,如何确定这些样例的类别标签将直接影响分类器的性能.基于少量的真实评论和大量的未标注评论,提出一种创新的 PU(positive and unlabeled)学习框架来识别虚假评论.首先,从无标注数据集中识别出少量可信度较高的负例.其次,通过整合 LDA(latent Dirichlet allocation)和 *K*-means,分别计算出多个代表性的正例和负例.接着,基于狄利克雷过程混合模型(Dirichlet process mixture model, DPMM),对所有间谍样例进行聚类,混合种群性和个体性策略来确定间谍样例的类别标签.最后,多核学习算法被用来训练最终的分类器.数值实验证实了所提算法的有效性,超过当前的基准.

关键词　虚假评论;全监督学习;PU 学习;狄利克雷过程混合模型;多核学习

中图法分类号　TP391

收稿日期:2013-11-14;**修回日期**:2014-08-12

基金项目:国家自然科学基金项目(61133012,61173062,61373108));国家哲学社会科学重大计划招标项目(11&ZD189)

电子商务的发展促进了在线用户评论数量的急剧增长,这些评论信息既可以帮助消费者作为参考来选择产品,又可以帮助商业群体及时获取消费者对其产品和服务的评价信息,以调整其产品和市场策略.因此,基于产品评论的情感分析和观点挖掘成为人工智能研究的热门话题[1-2].

情感分析和观点挖掘的研究工作主要包括情感分类、观点提取和特征抽取[3-4].这些研究有着共同前提:所采用的数据集(评论文本的集合)是真实可信的.由于观点信息可以引导消费者的购买行为,好评论会给商业组织或个体带来好名声和巨大经济收益,这促使了虚假评论的产生.通过前人的观察和总结[5-6],虚假评论可被分为以下 2 类:

1) 欺骗性评论.故意写下好评论促进产品销售,或故意写下差评论破坏产品名声,由此而产生的评论.

2) 破坏性评论.此类评论主要包括 3 个方面:①广告;②单独评价商标的评论,其内容不涉及所买的具体产品;③随机文本,不包含观点信息.

人们可以轻松识别出破坏性评论.而欺骗性评论具有隐藏性和多样性等特点,识别非常困难,因此欺骗性评论识别研究是一项艰巨而刻不容缓的任务.本文集中于欺骗性评论的识别.

通常情况下,虚假评论识别被认为是一个二分类问题[5-8],通过手工标注的数据集训练一个分类器,然后一个新的评论被预测为虚假或真实评论.但是由于虚假评论的隐藏性和多样性,同时现有的研究[6]表明,人类手工标注的评论数据集中必定存在着一定数量的误例,这些误例会影响分类器的生成能力.因此,简单使用传统的全监督分类算法来识别虚假评论并不合理.本文的研究动机如下:人类无法通过先验知识获取虚假评论集,但是却可以通过启发式的规则获取少量真实评论和大量的未标注评论,基于真实评论集和未标注评论集,能否建立一个准确的分类器用于识别虚假评论.

PU(positive and unlabeled)学习算法[9]可用于解决上述问题.不同于传统的全监督学习算法,PU 学习算法最主要特征是训练集中不含负例的情况下,分类器仍然可以获得准确性能.尽管已有的 PU 学习算法已被用于很多任务中[10-14],但是却尚未被用于识别虚假评论.另外,在应用 PU 学习算法时,数据集中的未标注评论集中肯定存在着一部分间谍样例(容易被误标注),设计有效策略来确定这些样例的类别标签是 PU 学习算法的关键.

本文提出基于 PU 学习算法来识别虚假评论.对于未标注评论集中的间谍样例,首先使用 DPMM 对其进行聚类[15],挖掘间谍样例的内部相关性.然后混合种群性和个体性 2 种策略确定间谍样例的类别标签,最后使用多核学习算法来训练最终的分类器.本文的主要贡献总结如下:

1) PU 学习算法首次被设计用于识别虚假评论.人类无法获取有效的虚假评论集,而 PU 学习算法只需少量的真实评论和大量的未标注评论就可建立合适的分类器.所以 PU 学习框架非常适用于这个任务.

2) 设计创新的策略来确定间谍样例的类别标签.在确定其标签时,使用 DPMM 算法对所有间谍样例聚类,然后同时考虑间谍样例的种群性和个体性来确定其标签.

3) 将多核学习算法应用于 PU 学习框架下.先前的 PU 学习算法使用单核 SVM 来训练最终的分类器.考虑到本文使用的数据集中样本较少,以及评论分布的多样性,使用多核学习将数据映射到更高维的特征空间来表达,提高最终的分类准确率.

1 相关研究工作

1.1 虚假评论识别

近 10 年来,研究者们在垃圾邮件[16-17]和垃圾网页[18]的识别研究上呈现大量工作,获得较好效果.最近,研究者们逐渐转向虚假评论识别研究.

Jindal 等人[5]发现虚假评论是广泛存在并且本质上不同于垃圾邮件和垃圾网页,他们利用产品评论数据,基于评论文本、评论者和产品的特征进行建模来区分复制观点(被认为是虚假评论)和非复制观点(被认为是真实评论).Wu 等人[19]提出使用基于产品的流行度排序是否被打乱来识别欺骗性评论.

Ott 等人[6]利用众包平台构造了虚假评论的黄金数据集.基于计算语言学和心理语言学的知识,他们采用传统的文本分类技术来识别虚假评论.本文将使用该数据集进行虚假评论识别研究.

Li 等人[7]从互联网的交易网站抓取一些产品评论数据,手工标注一个数据集,使用半监督的协同训练算法来识别虚假评论.但是,该方法使用的数据集为人工标注,Ott 等人在其研究中已经证实,人类无法通过先验知识有效识别虚假评论,手工标注的语料库中必将存在大量的误例.因此,采用人工标注数据集来进行虚假评论识别并不是最合理的方法.

Feng 等人[8]假设产品评论观点存在着先天的自然分布,并设计一系列实验来验证这个假设,发现评论文本上下文无关文法(context free grammar, CFG)的特征建模有助于提高检测性能.

不同于上述方法,考虑到虚假评论数据集构造的困难性,本文提出使用 PU 学习算法来识别虚假评论,下面简单介绍 PU 学习的相关算法.

1.2 PU 学习算法

根据 PU 算法对未标注数据集使用情况,将其分为以下 2 类.

1) 仅仅通过正例和未标注数据集中的部分样例来建立最终分类器[9,12]. 这类算法核心思想是:首先从未标注数据集中识别出可信度较高的负例,然后基于所有正例和可信负例,迭代使用期望最大化算法(expectation maximization, EM)算法或者支持向量机(support vector machine, SVM)建立最终分类器. 由于这些算法仅仅使用未标注数据集中的部分样例,而未考虑未标注数据集中剩余的大部分样例,因此训练出的分类器性能受到一定限制.

2) 使用正例和未标注数据集中的所有样例来建立最终分类器. 这类算法的关键是如何确定未标注数据集中样例的类别标签. Li 等人[11]成功将 PU 学习应用于传统算法不易处理的数据流环境下,提出基于相似度聚类的 PU 学习算法(PU learning by extracting likely positive and negative micro-cluster, LELC)用于数据流环境下的文档分类. Xiao 等人[13]提出基于相似度的 PU 学习算法(similarity based PU learning, SPUL),首先从未标注数据集中抽取一部分负例,计算剩余样例分属正例和负例的概率权重,然后重写包含上述概率权重的 SVM 优化函数和约束条件,最后训练出合适的 SVM 分类器.

不同于上述算法,本文在对未标注数据集中的剩余样例(间谍样例)进行处理时,首先使用无参贝叶斯模型 DPMM 聚类,捕捉间谍样例的内部联系. 然后混合种群性和个体性 2 种策略确定其类别标签,最后使用多核学习训练最终分类器.

2 PU 学习算法识别虚假评论

本节将详细介绍文中提出的 PU 学习算法. 首先,将虚假评论识别问题定义到 PU 学习框架中.

2.1 问题定义及符号标记

当前的互联网环境下,人们可以通过先验知识和启发式规则收集少量的真实评论,本文把真实评论构成的集合记为正例集合 P(称虚假评论为负例). 同时不难收集大量的未标注评论,记为 U. 基于数据集 P 和 U,本文提出一种混合种群性和个体性的 PU 学习(mixing population and individual nature PU learning, MPINPUL)算法,MPINPUL 算法分为以下 4 步:

Step1. 抽取可信负例;

Step2. 计算代表性样例;

Step3. 确定 U 中间谍样例的类别标签;

Step4. 建立最终分类器.

下面将对每一步进行详细的介绍.

2.2 抽取可信负例

由于数据集中只包含正例和未标注数据,因此,MPINPUL 算法的首要任务是抽取一些可信的负例. 先前工作主要有 2 种算法实施这一步,Spy-EM[9]和 Roc-SVM[12],本文整合这 2 种技术来抽取负例. 即分别用上述 2 种技术来抽取可信负例,最后取其交集作为本文的可信负例集合. 得到的可信负例存放在集合 RN 中. 未标注数据集 U 剩下的样例,即间谍样例,存放在集合 US 中.

2.3 计算代表性样例

第 1 阶段获得可信负例集合 RN,加上训练集中的正例集合 P,即可训练一个分类器. 但是,该分类器的性能不高,主要原因是忽略未标注数据集 U 中大量的剩余样例,即间谍样例集合 US,这些样例对提高分类器的性能有着重要作用. 为了确定间谍样例的标签,首先需要计算出能代表正例和负例的样例. 考虑到虚假评论和真实评论在语言结构和主题分布上的多样性,分别使用一个样例来代表虚假评论和真实评论并不合理,因此,这里提出使用 LDA(latent Dirichlet allocation)算法对 RN 进行聚类[20],然后计算出 2 个类别的多个代表性样例.

LDA 模型[20]认为每个文档可以表示为若干主题之间的分布,每个文档通过共同的主题相关联,是一个参数贝叶斯聚类模型,它可以很好捕捉文档内部间的关系. 在算法的第 2 阶段,首先利用 LDA 获得可信负例集合中每个样例在不同主题上的分布,接着使用 K-means 算法对可信负例进行聚类,使得可信负例中主题分布较一致的样例聚成一类. 最后通过使用 Rocchio 分类器分别为正例和负例计算出 10 个代表性样例,如算法 1 所示.

算法 1. 计算代表性样例算法.

输入:P 和 RN;

输出:$\boldsymbol{p}_k$ 和 $\boldsymbol{n}_k$,$k=1,\cdots,10$.

① 基于 LDA 算法,将 RN 聚类成 10 个子类 $RN_1, RN_2, \cdots, RN_{10}$;

② 将 P 和 RN 中的每个样例向量化为 $\boldsymbol{tf}\times\boldsymbol{idf}$;

③ FOR $k=1,\cdots,10$ DO

④ 通过式(1)计算 $\boldsymbol{p}_k$;

⑤ 通过式(2)计算 $\boldsymbol{n}_k$;

⑥ END FOR

$$\boldsymbol{p}_k = \alpha \frac{1}{|P|}\sum_{\boldsymbol{e}\in P}\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} - \beta\frac{1}{|RN_k|}\sum_{\boldsymbol{e}\in RN_k}\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|}; \tag{1}$$

$$\boldsymbol{n}_k = \alpha \frac{1}{|RN_k|}\sum_{\boldsymbol{e}\in RN_k}\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} - \beta\frac{1}{|P|}\sum_{\boldsymbol{e}\in P}\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|}. \tag{2}$$

算法 1 中:步骤①使用 LDA 将 RN 聚成 10 个子类. 步骤②中每个样例被向量化为 $\boldsymbol{v}=(q_1, q_2, \cdots, q_n)$,向量 $\boldsymbol{v}$ 的元素 q_i 通过一个单词 w_i 在文本中出现的频次 tf_i(term frequency)同该单词的逆文档频率 idf_i(inverse document frequency)的乘积来计算,即 $q_i=tf_i\times idf_i$. idf_i 的计算如下:

$$idf_i = \ln\frac{|D|}{df(w_i)}. \tag{3}$$

这里,$|D|$代表出现过单词 w_i 的评论的总数目. 从步骤③~⑥中,$\boldsymbol{p}_k$ 和 $\boldsymbol{n}_k$ 分别代表正例和负例的代表性样例,根据文献[11,13],α 和 β 分别被设置为 16 和 4.

2.4 确定间谍样例的类别标签

本文所提 PU 算法框架中的关键一步是确定间谍样例的类别标签,即令 $US=LP\cup LN$,其中 LP 存放 US 中标记为正例的间谍样例,LN 存放标记为负例的间谍样例.

算法 1 分别为正例和负例建立 10 个代表性样例,这里要利用这些代表性样例来确定每个间谍样例的类别标签. 由于算法 1 中采用的 Rocchio 技术是一个线性分类器,而虚假评论和真实评论的真实决策边界不一定是线性的,因此简单计算每个间谍样例同代表性样例的相似度来确定其类别标签将导致一定的错误. 本文提出首先使用 DPMM 对间谍样例聚类,然后混合间谍样例的种群性和个体性,一定程度上减少间谍样例的类别标注误差.

2.4.1 DPMM

基于 DP(Dirichlet process)的混合模型是近年来统计学习理论的研究热点[15,21-22],并且有着成功的应用[23-25],例如,Sun[23]提出对单个高斯过程进行扩展,实现了 IMGP(infinite mixtures of Gaussian processes)的变分估计,并将其成功用于交通流量预测. Hu 等人[24]把 HDP(hierarchical Dirichlet process)同 HMM(hidden Markov model)结合起来,成功用于视觉文档分析. DPMM 是一种基于狄利克雷过程混合模型的非参数贝叶斯聚类方法,它根据观测数据自动优化模型的结构,使模型的分布参数随着观测数据进行调整,具有天然的聚类特性. 本文使用 DPMM 对间谍样例进行聚类,无需事先设定类别个数,算法根据间谍样例集合,自动建立间谍样例间的关联性,从而聚成一类. 然后混合样例的种群性和个体性来确定其类别标签. 本文中,DPMM 将间谍样例集合 US 聚成不同的子类 $US_1, US_2, \cdots, US_m$.

2.4.2 样例的种群性

种群性的基本思想是相同子类中的样例应有很高可能性属于同样的类别. 本文设计 2 个算法:一个是子类种群性的计算;另一个是单个样例种群性的计算. 算法 2 展示子类的种群性,对于间谍样例的每个子类,首先计算出子类中每个样例的最相似代表性样例,确定每个样例的暂时类别标签. 然后使用大多数投票法则决定整个子类的类别标签. 最终的子类类别标签作为子类中每个样例的类别标签.

算法 2. 子类种群性的计算算法.

输入:$US_i, i=1,2,\cdots,m$;

输出:LP_i 和 LN_i.

① $LP_i=\varnothing, LN_i=\varnothing, p\_vote=0, n\_vote=0$;

② FOR 每个样例 $\boldsymbol{e}\in US_i$ DO

③ IF $\max\limits_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{p}_i) > \max\limits_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{n}_i)$

④ THEN p_vote++;

⑤ ELSE n_vote++;

⑥ END IF

⑦ END FOR

⑧ IF $p\_vote>n\_vote$

⑨ THEN $LP_i=LP_i\cup US_i$;

⑩ ELSE $LN_i=LN_i\cup US_i$;

⑪ END IF

算法 2 中,

$$sim(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|}. \tag{4}$$

$\max\limits_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{p}_i)$代表 $sim(\boldsymbol{e},\boldsymbol{p}_i)$在 10 种情况下的最大值,算法 2 可计算出子类 US_i 的类别标签.

对于单个样例的种群性,首先确定单个样例所处子类,然后用子类类别标签作为该样例的类别标签,如算法 3 所示.

算法 3. 单个样例种群性的计算算法.

输入:样例 $\boldsymbol{e}$;

输出:LP_i 和 LN_i.

① $LP_i=\varnothing, LN_i=\varnothing, p\_vote=0, n\_vote=0$;

② 确定样例 $\boldsymbol{e}$ 所在的子类 US_i;

③ FOR 每个样例 $\boldsymbol{e}\in US_i$ DO

④ IF $\max\limits_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{p}_i) > \max\limits_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{n}_i)$

⑤ THEN p_vote++;

⑥ ELSE n_vote++;

⑦ END IF

⑧ END FOR

⑨ IF $p\_vote > n\_vote$

⑩ THEN $LP_i=LP_i\cup\{\boldsymbol{e}\}$;

⑪ ELSE $LN_i=LN_i\cup\{\boldsymbol{e}\}$;

⑫ END IF

算法 2 确定子类的类别标签,该算法不仅考虑每个样例同正例和负例代表性样例的关系,同时也考虑子类中样例间的内在关系.但该算法有个缺点,即当子类中实际的正例和负例数目比较接近时,算法效果较差.图 1 呈现 DPMM 对间谍样例集合 US 的部分聚类结果 Micro-C1, Micro-C2, Micro-C3, Micro-C4,假设黑实线代表虚假评论和真实评论的边界.根据算法 2,Micro-C1 和 Micro-C2 可以很好确定子类的类别标签,以此来决定每个样例的类别标签.Micro-C3 中大部分是正例,很小一部分是负例,这种情况下,子类中的每个样例可根据子类的类别标签,即正例来确定.这里存在少量的类别标注错误是可容忍的.但对于 Micro-C4,正例和负例数目比较接近,如果采用子类的类别标签来决定每个样例的类别标签将会产生一定量错误.因此,在这种情况下,本文提出用混合样例的个体性来弥补群体性的不足.

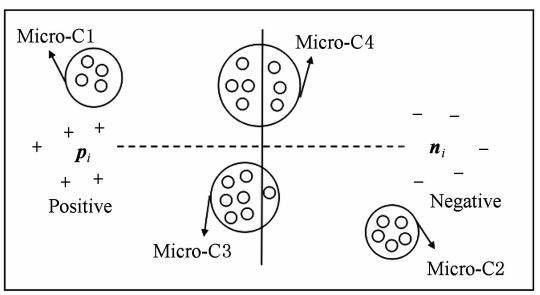

Fig. 1 Illustration of population nature.

图 1 子类的种群性

2.4.3 样例的个体性

个体性的基本思想是:应充分考虑单个样例同所有正例和负例的代表性样例间的相似关系,忽略样例所处子类的群体间关系.具体地,首先计算单个样例同所有代表性样例的相似度,然后使用式(5)和式(6)计算出样例属于正例和负例的类别概率:

$$proba\_positive(\boldsymbol{e})=\frac{\sum\limits_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{p}_i)}{\sum\limits_{i=1}^{10}(sim(\boldsymbol{e},\boldsymbol{p}_i)+sim(\boldsymbol{e},\boldsymbol{n}_i))};\tag{5}$$

$$proba\_negative(\boldsymbol{e})=\frac{\sum\limits_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{n}_i)}{\sum\limits_{i=1}^{10}(sim(\boldsymbol{e},\boldsymbol{p}_i)+sim(\boldsymbol{e},\boldsymbol{n}_i))}.\tag{6}$$

以 Micro-C3 为例,该子类中的样本标签如果按照子类的标签来确定,将产生一定的误差,在这种情况下,可考虑使用样例的个体性计算每个样例的类别标签,样例的个体性如算法 4 所示.

算法 4. 样例个体性的计算算法.

输入:$US_i, i=1,2,\cdots,m$;

输出:LP_i 和 LN_i.

① $LP_i=\varnothing, LN_i=\varnothing$;

② FOR 每个样例 $\boldsymbol{e}\in US_i$ DO

③ IF $proba\_positive(\boldsymbol{e}) > proba\_negative(\boldsymbol{e})$

④ THEN $LP_i=LP_i\cup\{\boldsymbol{e}\}$;

⑤ ELSE $LN_i=LN_i\cup\{\boldsymbol{e}\}$;

⑥ END IF

⑦ END FOR

算法 4 可较准确地计算出子类 Micro-C3 中的样例的类别标签,从而避免大量的类别标注错误.

2.4.4 算法的 MPINPUL-1

根据 2.4.1 节至 2.4.3 节的讨论,为了更准确确定间谍样例的类别标签,可以混合群体性和个体性来设计算法确定间谍样例的类别标签.第 1 种思路算法 MPINPUL-1 是先考虑子类的群体性,当群体性不满足一定阈值时,为了减少类别标注误差,采用个体性来分别确定子类中每个样例的类别标签.即先群体性,后个体性.具体 MPINPUL-1 算法如下:

算法 5. MPINPUL-1.

输入:US, p_k 和 n_k, $k=1,2,\cdots,10$;

输出:LP 和 LN.

① $LP=\varnothing, LN=\varnothing$;

② 基于 DPMM,将 US 聚成 n 个子类 US_1, $US_2,\cdots,US_n$;

③ FOR 每个子类 US_i DO

④ $p\_number=0, n\_number=0$;

⑤ FOR 每个样例 $e\in US_i$ DO

⑥ IF $\max_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{p}_i) > \max_{i=1}^{10} sim(\boldsymbol{e},\boldsymbol{n}_i)$

⑦ THEN p_number++;

⑧ ELSE n_number++;

⑨ END IF

⑩ END FOR

⑪ IF((|p_number-n_number|)/|US_i|)<s_1

⑫ THEN 执行算法 4; /* 个体性 */

⑬ ELSE 执行算法 2; /* 群体性 */

⑭ END IF

⑮ $LP=LP\bigcup LP_i$;

⑯ $LN=LN\bigcup LN_i$;

⑰ END FOR

算法 5 中,参数 s_1 的选择是至关重要的,具体实验时,根据先验知识设置几个阈值,选择算法效果最好的参数.

2.4.5 算法 MPINPUL-2

部分虚假评论是由拥有真实购物经历和有过真实评论经历的人所写,这部分评论在语言结构和情感极性上同真实评论比较接近,不是很好识别. 基于此,本文设计算法 MPINPUL-2 来确定间谍样例的类别标签. 算法的基本思路是:计算单个样例属于正例、负例的类别概率,当 2 个概率比较接近时,使用该样例所处子类的类别标签来确定该样例的类别标签. 即先个体性、后群体性. 这是混合群体性和个体性 PU 学习算法 MPINPUL 的第 2 种思路,具体算法如下:

算法 6. MPINPUL-2.

输入:US, p_k 和 n_k, $k=1,2,\cdots,10$;

输出:LP 和 LN.

① $LP=\varnothing$, $LN=\varnothing$;

② 基于 DPMM,将 US 聚成 n 个子类 US_1, US_2, …, US_n;

③ FOR 每个子类 US_i DO

④ FOR 每个样例 $\boldsymbol{e}\in US_i$ DO

⑤ IF (|$proba\_positive(\boldsymbol{e})$-$proba\_negative(\boldsymbol{e})$|)<$s_2$

⑥ THEN 执行算法 3; /* 群体性 */

⑦ ELSE 执行算法 4; /* 个体性 */

⑧ END IF

⑨ END FOR

⑩ $LP=LP\bigcup LP_i$;

⑪ $LN=LN\bigcup LN_i$;

⑫ END FOR

跟算法 MPINPUL-1 中的 s_1 一样,算法 MPINPUL-2 的参数 s_2 的选择也至关重要,我们将在本文的 4.2 节对其进行讨论.

2.5 建立最终分类器

基于正例集合 P 和可信负例集合 RN,以及本文所提算法第 3 阶段计算出的正例集合 LP,负例集合 LN,即可训练最终的 SVM 分类器. 由于本文使用的数据集所包含的样例较少,为了提高分类器的性能,使用多核学习[26-27]来训练最终的分类器.

传统的 SVM 采用一个核函数,不足以解决数据集分布不平坦、数据来源异构下的复杂问题. 多核学习算法使用多个核函数将输入空间变换到高维的特征空间,转化为求解凸优化问题,并求解全局最优解. 文献[28-32]论述的理论和应用已经证明多核代替单核能增强决策函数的可解释性,并能获得比单核模型和单核及其组合模型更好的性能. 在本文所提的 PU 框架下,分别使用 2 种主流多核学习算法 SILP[29] 和 SimpleMKL[30] 来进行实验,其中,SILP 通过 Matlab 的 LPSOLVE toolbox 求解,SimpleMKL 由文献[30]给出的 SVM-KM toolbox 求解. 建立最终分类器的算法如下:

算法 7. 建立最终分类器.

输入:P, RN, LP 和 LN;

输出:分类器 F_{MK}.

① $P=P\bigcup LP$;

② $N=RN\bigcup LN$;

③ 基于正例集合 P 和负例集合 N,使用多核学习算法 SILP 或 SimpleMKL 训练出最优的多核分类器 F_{MK}.

3 数值实验

本文目标是判断 PU 学习算法能否有效用于虚假评论识别. 实验之前介绍本文使用的数据集. 首先介绍数据集的收集方式,接着介绍如何将数据集划分成本文使用的正例集合 P 和未标注数据集 U. 然后讨论算法的评价准则. 最后简单呈现了人类在该数据集上的识别性能.

3.1 数据集

Ott 等人[6]利用众包平台获取黄金数据集,这是唯一公开可用的数据集. 本文采用该数据集进行虚假评论识别. 数据集中的评论是针对芝加哥地区的酒店. 下面介绍数据集中的虚假评论和真实评论.

3.1.1 虚假评论

众包服务可以实现大规模数据收集和数据标注等服务.由于虚假评论的隐蔽性和多样性,人类无法从现有的评论中标注出虚假评论,却可以人工来构造一部分虚假评论.Ott 等人利用众包平台,定义了 400 个任务,每个任务的设定是:假如你是酒店市场部门的工作人员,让你写一条有利于酒店发现的正面评论.每个任务的报酬是一美元.基于此,他们收集 400 条虚假评论.

3.1.2 真实评论

真实评论的收集来自于 TripAdvisor①,针对芝加哥同样酒店的 6 977 条评论.对这些评论进行以下处理:

1) 删除 3130 条非 5 星级评论;

2) 删除 41 条非英语评论;

3) 删除少于 75 个字符的评论;

4) 删除作者是第 1 次发表的评论.

为了平衡虚假评论的数目,以及同虚假评论长度的分布保持一致.他们在剩余的 2 124 条评论中,选择出 400 条评论作为真实评论.最终,总共 800 条评论构成了黄金数据集.

3.2 实验设置和评价

本文采用十折交叉验证:数据集被划分成 10 等份,9 份用来作为训练集,第 10 份被用来作为测试集.算法运行 10 次,每次选择不同的测试集,最后取 10 次平均值作为最后结果.算法每次运行过程中,测试集每次包含 80 个评论,40 个真实评论和 40 个虚假评论.对于训练集来说,里面包含 360 个真实评论和 360 个虚假评论.由于本文旨在使用 PU 学习算法,特进行以下设定:取训练集中真实评论的 20%,作为正例集合 P,剩余的所有真实评论和虚假评论作为未标注数据集 U.因此,算法的一次运行过程中,训练集包含 720 个评论,其中正例集合 P 包含 72 个评论,未标注数据集 U 包含 648 个评论,而测试集包含 80 个评论.

由于本文使用的数据集实际上具有类别标签,所以采用准确率评价最终分类器的性能.

3.3 人工性能

评估人工性能基于以下 2 方面原因:1)如果人工性能很低的话,说明标注语料的困难性,从而说明使用 PU 学习算法来解决这个任务的必要性;2)人工性能作为一个基准,可以跟本文提出的算法相比较.

为了评估人工性能,求助于 3 个计算机专业的本科生.他们被要求在测试集上(包含 40 个真实评论和 40 个虚假评论)作判断.这里使用 2 种元判断方式:MAJORITY 和 SKEPTIC,MAJORITY 表示 3 人中至少有 2 人判定是虚假评论才确定该评论属于虚假评论,而 SKEPTIC 表示 3 人中只要有 1 人认为是虚假评论就判定该评论属于虚假评论.从表 1 的统计结果中显示,人工性能是非常低的.接着对标注一致性进行统计分析,3 人中两两标注一致的 Fleiss 卡帕值为 0.09,Landis 和 Koch 在文献[33]中证实这个值暗示标注者之间轻微的一致.而 JUDGE 1 和 JUDGE 3 的 Cohen 卡帕值为 0.11,该值低于可接受的一致标注水平.基于上述分析,可以认为标注虚假评论数据集是非常困难的.因此本文提出将虚假评论识别问题定义在 PU 学习框架下是非常合适的.

Table 1 Human Performance

表 1 人工性能 %

Methods	Accurancy	Recall	F
JUDGE1	56.8	54.9	55.8
JUDGE2	53.2	52.2	52.7
JUDGE3	59.2	59.1	59.1
MAJORITY	58.1	49.6	57.6
SKEPTIC	60.2	59.8	60.1

4 实验结果和分析

为了验证本文所提算法的有效性,本文额外进行以下工作.首先,实现 2 种主流 PU 学习算法 LELC[11] 和 SPUL[13].其次,将 2 种主流的虚假评论识别算法放在提出的 PU 框架的第 4 个阶段进行实验.

4.1 实验结果

首先讨论本文所提 PU 框架同先前 PU 算法的比较.固定 $s_1=0.1$,$s_2=0.1$,表 2 给出不同 PU 算法在数据集上的识别性能.由表 2 可知,本文设计 PU 框架的最好结果是 MPINPUL-2 和 SimpleMKL 的组合,其可获得 83.21%的准确率,同时实验先前已有的 PU 框架,LELC 和 SPUL,其分别获得 81.12% 和 81.89%的准确率,这说明本文所提的 PU 框架要好于先前的算法.同时,本文所提的 MPINPUL-1 和 MPINPUL-2 同多核学习 SILP 组合也获得不错的效果.例如,MPINPUL-2 和 SILP 的组合,获得高达 82.96%的准确率,也超过 LELC 和 SPUL,这充

① http://www.tripadvisor.co

分说明了本文所设计的 PU 学习框架的有效性.

Table 2 Overall Accuracy on Different PU Learning

表 2 已有 PU 算法的比较 %

Methods	Accuracy
LELC	81.12
SPUL	81.89
MPINPUL-1+SILP	82.84
MPINPUL-2+SILP	82.96
MPINPUL-1+SimpleMKL	82.85
MPINPUL-2+SimpleMKL	83.21

为了验证本文所提算法的有效性,在本文所提的 PU 框架下,在建立最终分类器的阶段采用了 2 种主流的虚假评论识别算法来进行实验,这 2 种算法分别由 Ott 等人和 Feng 等人提出. 如表 3 所示,2 种主流的虚假评论识别算法分别获得 80.13%和 80.72%的准确率,而使用多核学习算法训练最终的分类器平均可得到 83.1%左右的分类准确率,比先前的虚假评论算法提高了 3%左右. 主要原因是本文使用了使用多核学习算法,能将数据映射到更高维的空间进行区分,因此获取更优的效果.

Table 3 Performance on Different Algorithms

表 3 与虚假评论识别算法的比较 %

Methods	Accuracy
MPINPUL-1+Ott et al.	80.13
MPINPUL-2+Ott et al.	80.72
MPINPUL-1+Fen et al.	80.43
MPINPUL-2+Feng et al.	80.95
MPINPUL-2+SILP	82.96
MPINPUL-2+SimpleMKL	83.21

本文所提 PU 学习算法获得较好的性能,主要原因归结于以下 3 个方面:1)无参数贝叶斯模型 DPMM 的使用,捕获了样例的内部深层次的联系,获取了更好的聚类结果;2)混合群体性和个体性的策略,使得间谍样例标签的确定更加准确;3)多核学习算法的使用,将特征映射到高维的空间进行区分.

4.2 参数敏感性

在上述实验中,固定 $s_1=0.1, s_2=0.1$,这里将选择不同的参数 s_1 和 s_2 进行实验,考察 2 个参数对算法性能的影响. 基于算法 MPINPUL-1,结合 SimpleMKL,选择区间[0.1,0.5]内不同 s_1 来进行实验,结果如图 2 所示,可知,当 s_1 位于区间[0.1,0.2]时,算法 MPINPUL-1 可获得较好的效果,当$s_1>0.2$ 时,随着 s_1 的增大,算法的性能近似呈线性下降.

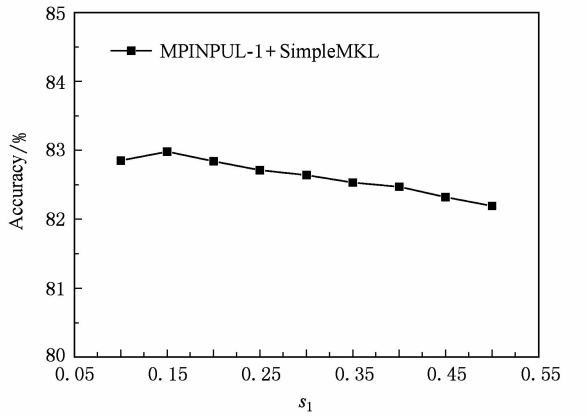

Fig. 2 Performance on different s_1.

图 2 算法 MPINPUL-1 随 s_1 值的变化情况

类似于 MPINPUL-1,本文也实验了不同 s_2 对算法 MPINPUL-2 的性能变化情况(结合 SimpleMKL). 实验结果如图 3 所示. 当 s_2 位于区间[0.1,0.2]时,算法 MPINPUL-2 可获得最好性能,当 $s_2>0.2$,随着 s_2 的增大,算法 MPINPUL-2 的性能大致呈线性下降.

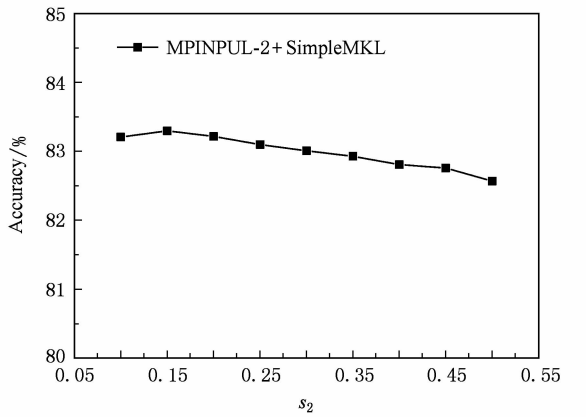

Fig. 3 Performance on different s_2.

图 3 算法 MPINPUL-2 随 s_2 值的变化情况

5 总 结

本文提出使用 PU 学习算法来识别虚假评论. 在确定间谍样例的类别标签时,首先采用无参数贝叶斯模型 DPMM 对其进行聚类,然后混合种群性和个体性 2 种策略,来确定间谍样例的类别标签,最后使用多核学习算法训练最终分类器. 数值实验证实了所提 PU 学习算法可以有效用于虚假评论识别.

未来的工作考虑将本文提出的 PU 算法应用于大规模的数据集中.

参 考 文 献

[1] Lin Zheng, Tan Songbo, Cheng Xueqi. Sentiment classification analysis based on extraction of sentiment key sentence [J]. Journal of Computer Research and Development, 2012, 49(11): 2376-2382 (in Chinese)
(林政,谭松波,程学旗. 基于情感关键句抽取的情感分类研究[J]. 计算机研究与发展, 2012, 49(11): 2376-2382)

[2] Li Suke, Jiang Yanbing. Semi-supervised sentiment classification based on sentiment feature clustering [J]. Journal of Computer Research and Development, 2013, 50(12): 2570-2577 (in Chinese)
(李素科,蒋严冰. 基于情感特征聚类的半监督情感分类[J]. 计算机研究与发展, 2013, 50(12): 2570-2577)

[3] Murthy G, Liu B. Mining opinions in comparative sentences [C] //Proc of the 22nd Int Conf on Computational Linguistics (COLOING'08). New York: ACM, 2008: 241-248

[4] Hu Mingqing, Liu Bing. Mining opinion features in customer reviews [C] //Proc of the 19th National Conf on Artificial Intelligence(AAAI'04). San Jose: American Association for Artificial Intelligence, 2004: 775-760

[5] Nitin J, Liu B. Opinion spam and analysis [C] //Proc of the 1st ACM Int Conf on Web Search and Data Mining (WSDM'08). New York: ACM, 2008: 137-142

[6] Myle O, Choi Y L, Claire C, et al. Finding deceptive opinion spam by any stretch of the imagination [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL'11). Stroudsburg, PA: Association for Computational Linguistics, 2011: 309-319

[7] Li Fangtao, Huang Minlie, Yang Yi, et al. Learning to identify review spam [C] //Proc of the 22nd Int Joint Conf on Artificial Intelligence(IJCAI'2011). San Francisco: Morgan Kaufmann, 2011: 2488-2493

[8] Feng S, Ritwik B, Choi Y J. Syntactic stylometry for deception detection [C] //Proc of the 50th Annual Meeting of the Association for Computational Linguistics (ACL'12). Stroudsburg, PA: Association for Computational Linguistics, 2012: 171-175

[9] Liu B, Wee S L, Philip S Y, et al. Partially supervised classification of text documents [C] //Proc of the 19th Int Conf on Machine Learning(ICML'02). New York: ACM, 2002: 387-394

[10] Charles E, Keith N. Learning classifiers from only positive and unlabeled data [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Ming(SIGKDD'08). New York: ACM, 2008: 213-220

[11] Li X L, Philip S Y, Liu B, et al. Positive unlabeled learning for data stream classification [C] //Proc of the 9th SIAM Int Conf on Data Ming(SDM'09). Philadelphia, PA: SIAM, 2009: 257-268

[12] Liu B, Yang D, Li X L, et al. Building text classifiers using positive and unlabeled examples [C] //Proc of the 3rd IEEE Int Conf on Data Ming(ICDM'03). Piscataway, NJ: IEEE, 2003: 179-182

[13] Xiao Yanshan, Liu Bing, Yin Jie, et al. Similarity-based approach for positive and unlabeled learning [C] //Proc of the 22nd Int Joint Conf on Artificial Intelligence(IJCAI'11). San Francisco: Morgan Kaufmann, 2011: 1577-1582

[14] Dell Z. A simple probabilistic approach to learning from positive and unlabeled examples [C] //Proc of the 5th Annual UK Workshop on Computational Intelligence. Piscataway, NJ: IEEE, 2005: 83-87

[15] Yee W T, Michael I J, Matthew J B, et al. Sharing clusters among related groups: Hierarchical dirichlet processes [C] //Proc of Advances in Neural Information Processing Systems (NIPS'04). Cambridge, MA: MIT Press, 2004: 1385-1392

[16] Harris D, Wu D H, Vapnik N. Support vector machines for spam categorization [J]. IEEE Trans on Neural Networks, 1999, 10(5): 1048-1054

[17] Zoltan G, Hector G M, Jan P. Combating web spam web with trustrank [C] //Proc of the 30th Int Conf on Very Large Data Bases(VLDB'04). New York: ACM, 2004: 576-587

[18] Alexandros N, Marc N, Dennis F. Detecting spam web pages through content analysis [C] //Proc of the 15th Int Conf on World Wide Web (WWW'06). Berlin: Springer, 2006: 83-92

[19] Wu F, Bernardo A H. Opinion formation under costly express [J]. ACM Trans on Intelligence System Technology, 2010, 1(1): 1-13

[20] David M B, Andrew Y N, Michael I J. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[21] Neal R M. Markov chain sampling methods for Dirichlet process mixture models [J]. Journal of Computational and Graphical Statistics, 2000, 9(2): 249-265

[22] Yee W T, Michael J. Hierarchical bayesian nonparametric models with applications [M]. Bayesian Nonparametric: Principles and Practice. Cambridge, UK: Cambridge University Press, 2010

[23] Sun Shiliang. Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction [J]. IEEE Trans on Intelligent Transportation Systems, 2011, 12(2): 466-475

[24] Hu Weiming, Tian Guodong, Li Xi, et al. An improved hierarchical dirichlet process-hidden markov model and its application to trajectory modeling and retrieval [J]. International Journal of Computer Vision, 2013, 105(3): 246-268

[25] Huang Ruizhang, Yu Guan, Wang Zhaojun, et al. Dirichlet process mixture model for document clustering with feature partition [J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25(8): 1748-1759

[26] Yang Jingjing, Li Yuanning, Tian Yonghong, et al. Group-sensitive multiple kernel learning for object categorization [C] //Proc of the 12th Int Conf on Computer Vision (ICCV'09). Piscataway, NJ: IEEE, 2009: 436-443

[27] Kumar A, Sminchisescu C. Support kernel machines for object recognition [C] //Proc of the 11th Int Conf on Computer Vision(ICCV'07). Piscataway, NJ: IEEE, 2007: 1-8

[28] Li Jinbo, Sun Shiliang. Nonlinear combination of multiple kernels for support vector machines [C] //Proc of the 20th Int Conf on Pattern Recognition(ICPR'10). Piscataway, NJ: IEEE, 2010: 2889-2892

[29] Gert L, Nello C, Peter B, et al. Learning the kernel matrix with semi-definite programming [C] Journal of Machine Learning Research, 2004, 5: 27-72

[30] Alain R, Francis R B, Stephane C, et al. Simple MKL [J]. Journal of Machine Learning Research, 2008, 9: 2491-2521

[31] Sun Tao, Jiao Licheng, Liu Fang, et al. Selective multiple kernel learning for classification with ensemble strategy [J]. Pattern Recognition, 2013, 46(11): 3081-3090

[32] Cortes C, Mohri M, Rostamizadeh A. Multi-class classification with maximum margin multiple kernel [C] //Proc of the 30th Int Conf on Machine Learning (ICML'13). New York: ACM, 2013: 22-31

[33] Richard L, Gary G K. The measurement of observer agreement for categorical data [J]. Biometrics, 1977, 33 (1): 159-174

Ren Yafeng, born in 1986. PhD candidate at the School of Computer, Wuhan University. Student member of China Computer Federation. His research interests include natural language processing, machine learning and data mining, etc.

Ji Donghong, born in 1967. Received his PhD degree in the School of Computer from Wuhan University, China in 1995. Currently professor and PhD supervisor at the School of Computer, Wuhan University. His research interests include natural language processing, machine learning and data mining, etc(dhji@whu.edu.cn).

Zhang Hongbin, born in 1979. PhD candidate at the School of Computer, Wuhan University. Associate professor at the School of Software, East China Jiaotong University. His research interests include natural language processing and computer vision.

Yin Lan, born in 1979. PhD candidate at the School of Computer, Wuhan University. Associate professor of Guizhou Normal University. Her research interests include natural language processing and machine learning, etc.