生活消费平台虚假评论识别模型的研究

李 晶[1],吴国仕[1],谢 菲[2],姚 旭[1],齐佳音[3],孙鹏飞[1]
(1. 北京邮电大学软件学院,北京 100876; 2. 新华社通信技术局,北京 100803;
3. 北京邮电大学经济管理学院,北京 100876)

摘 要: 生活消费平台已成为人们获取商家信息、反馈服务或产品质量的重要平台. 虚假评论作为一种夸大或诽谤目标商家口碑的商业行为在生活消费平台很普遍,具有很强的危害性. 本文对某网站的真实评论展开虚假评论研究,深入分析研究虚假评论的特征,从"可信度"的角度出发,提出用户及商家可信度模型. 利用评论人的行为特征、商家的特征和评论文本的特征构建了虚假评论识别模型,经测试该模型达到了一个良好的识别效果.

关键词: 机器学习; 虚假评论识别; 可信度模型
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2016)12-2855-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.12.07

Research of Fraud Review Detection Model on O2O Platform

LI Jing[1], WU Guo-shi[1], XIE Fei[2], YAO Xu[1], QI Jia-yin[3], SUN Peng-fei[1]
(1. *School of Software Engineering, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *Communication and Technical Bureau, Xinhua News Agency, Beijing* 100803, *China*;
3. *School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: Living-consumption platform has become a very important platform for customers to extract information of businesses, and view or submit comments on the quality of services or products. It is common that fake reviews, as a commercial activity, are used to exaggerate or damage the reputation of a target business, which is extremely harmful. This paper chose an O2O (Online To Offline) platform, from which reviews are derived, to study fake reviews. With an in-depth study on features of fake reviews, it raised the user-credibility and shop-credibility evaluation model respectively from the credibility perspective. Based on features of reviewers, businesses, and review texts, it established a fake review identification model, and through testing this model showed excellent performance in identification.

Key words: machine learning; fraud review detection; credibility model

1 引言

生活消费平台作为一种新兴的互联网服务在近几年取得了巨大的发展. 某些组织或个人在利益的驱动下,利用网络信息监管的缺失,弄虚作假,制造虚假评论误导用户. 因此,对生活消费平台虚假评论进行研究并加以识别成为人们关注的技术热点.

评论可信度是评论人提供的信息被认可的程度. 国内关于商品评论的最新研究提出了基于文本内容的商品评论可信度测评模型[1]. 国外的研究发现匿名评论的情感倾向会对可信度产生较大影响[2];体验型产品中情感倾向对可信度几乎无影响[3];评论与评分及评论间的一致性越高,评论可信度越高[4~7]等. 可见,评论可信度受多种因素影响. 研究对象及特征组合的不同会影响结论的一致性.

虚假评论的特征主要从评论内容和评论人两个角度来考虑. 从评论内容的角度来分析,许多研究采用了词性和 n 元文法. Ott M[8]等利用一元、二元文法结合心理学构建的 80 个情感特征关键词获得了 90% 的查准率. 但 Li F T[9]等的研究结论显示,利用情感来分辨欺骗型评论的效果并不显著,因为在刻意虚构的评论中,这样的情感特征并不明显. 研究表明,单纯从评论的文本内容特征展开识别,对欺骗型评论的识别效果并不理想. 评论人的特征反应了评论撰写者的个人信用和

收稿日期:2015-01-31;修回日期:2015-05-20;责任编辑:梅志强
基金项目:国家 973 重点基础研究发展计划(No. 2013CB329604);国家自然科学基金(No. 71231002);北京市自然科学基金(No. 9122018);教育部博士点基金(No. 20120005110015);新华社 713 实验室技术研究项目——大数据与智能信息处理课题

行为特征,通过识别评论人的特征来识别其发表言论的特征具有一定的研究意义. Mukherjee A[10] 等针对 Yelp 的真实数据,比较了基于评论特征和基于评论人特征的虚假评论识别效果,发现后者的识别效果更好. Li F T[9] 也指出评论人的行为特征是评论内容特征的重要补充. 评论和评论人特征的抽取又和特定的领域有一定的相关性. 目前,国外对虚假评论的研究涉及到了旅馆、图书、音乐、餐馆等领域. 国内的研究数据对象主要集中在图书和数码产品领域,对服务行业还未涉及.

2 数据描述

本文数据取自某网站 2012 年 11 月到 2013 年 10 月的已标注数据子集,经清洗后共包含 331,415 条评论,238,186 评论者与 109,376 商家.

商家的评论数据包含 7 个基础部分:〈评论类型〉〈评论 ID〉〈商家 ID〉〈评论者 ID〉〈提交日期〉〈星级〉〈评论文本〈还有三项专项评分:〈口味评分〉〈环境评分〉〈服务评分〉

经测试,评论的数量与评论人的数量之间遵循幂律分布,即大量的用户只写了少数评论,少数用户却写了大量评论. 评论的数量与商家的数量之间同样遵循幂律分布,即大量的产品只获得很少评论,少数产品获得了大量的评论.

3 虚假评论识别模型研究

3.1 建模准则

依据前文论述评论与商家、用户呈现一对多的关系,对于许多商家或评论人甚至仅仅对应一条评论. 因此,商家、用户对于判断虚假评论具有重要的作用. 然而,用户与商家对于评论可信度的影响程度不同. 如果用户为虚假用户,则其所有评论都极有可能成为垃圾评论;而商家可能会偶尔出现为提高知名度雇佣水军刷评论的欺骗行为,但不一定所有的评论都是垃圾评论. 用户相比于商家与评论可信度具有较强的关系. 经计算,用户可信度与评论真实性的协方差为 0.58,而商家可信度与评论真实性的协方差为 0.47,因此用户可信度相比于商家可信度与评论可信度更相关,与常识相符.

在识别虚假评论时,采用分层的探测机制,首先识别虚假评论中呈现出群体特征的水军用户;之后分别构建用户、商家可信度模型;最后利用评论对应的商家可信度、用户可信度与其他特征建模识别虚假评论,模型结构如图 1 所示.

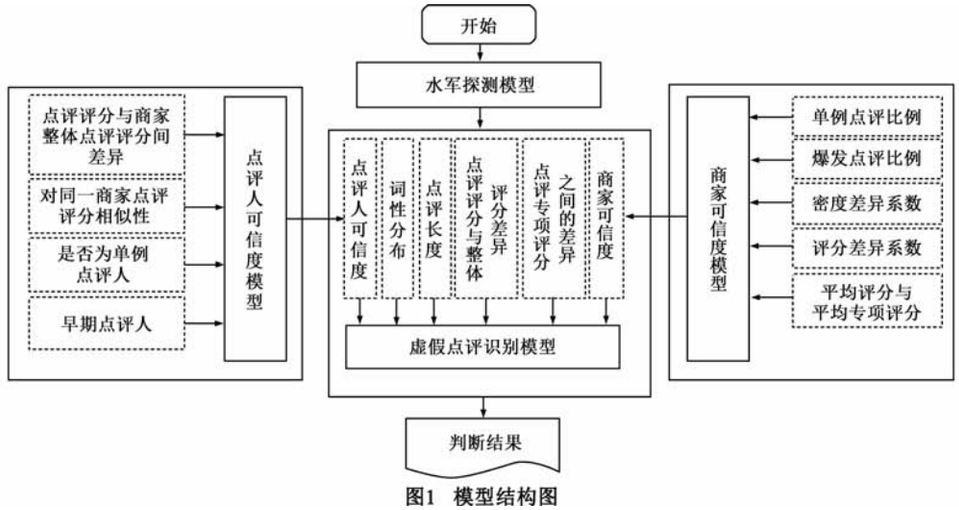

图1 模型结构图

3.2 水军探测

虚假评论人常常出现群体性活动的特征,将这种具有群体性特征的评论人称为水军. 频繁项集挖掘的 eclat 算法[11] 是探测水军的理想算法. 经测试,被探测出的包含在频繁项集内的 18 个用户中,有 15 个为虚假评论人,虚假率为 83.3%,因此对本数据集可直接将挖掘出的频繁项集作为垃圾评论人.

3.3 评论人与商家可信度模型

3.3.1 评论人可信度模型

为将评论人及商家可信度量化,本文采用逻辑回归模型[12].

评论人可信度模型的可表示为:

$$P_a = h_\theta(x_a) = g(\theta_a^t x_a) = \frac{1}{1+e^{-\theta_a^t x_a}} \tag{1}$$

$$\theta_a^t x_a = \theta_{1a}x_{1a} + \theta_{2a}x_{2a} + \cdots + \theta_{na}x_{na} \tag{2}$$

其中,x_{ka} 与 θ_{ka} 分别为评论人 a 的第 k 个特征与第 k 个特征的权重. 下面对评论人的不同特征做详细说明.

(1)评论评分与商家整体评论评分间差异

每个商家具有一个根据对其所有的评论评分计算出的整体评分. 从概率上讲,一个真实、理性的评论人打出

的分数应该和该整体评分差异不大;而虚假评论人,往往倾向给出与整体评分差异较大(即较高或较低)的评分. 通过计算用户的所有评论评分和相关商家的整体评分之间的差异性大小,可以将该特征进行量化.

以 RD_a 表示该特征为:

$$RD_a = \text{avg}_{p \in P_a} \frac{|r_{ap} - \overline{r_{ap}}|}{40} \tag{3}$$

其中,r_{ap} 指评论人 a 对商家 $p \in P_a$ 给出的评论评分,$\overline{r_{ap}}$ 指该商家的整体评分,为了归一化,本文在此处将二者差的绝对值除以 40(评分的正常取值范围为 10 ~ 50). 取该归一化差异值的平均值作为评论人的评论评分与商家整体评论评分的差异.

(2)对同一商家评论评分相似性

大多数正常情况下,评论人仅对某个商家发表至多一次评论. 少数正常情况下,评论人会发表一次以上的评论. 再次发表评论的评分应当与第一次评论的评分有一定的差异性. 然而,一些虚假评论人会多次评论同一商家,并且评分会比较接近,或完全相同. 通过量化该特征可以将其集成入用户可信度模型中,该特征包括两个考虑因素,一是用户对同一商家发表过多少次评论,二是不同评论间差异的大小. 以 C_a 来表示该特征为:

$$C_a = \sum_{p \in P_a} n_{ap}(1 - \text{CV}_{ap}) \tag{4}$$

其中,n_{ap} 为评论人 a 对商家 $p \in P_a$ 所有评论的数目,CV_{ap} 是 a 对商家 $p \in P_a$ 所有评论评分的差异系数(Coefficient of Variation),其计算方法为所有评论评分的标准差比平均值,即:

$$C_a = \sum_{p \in P_a} n_{ap}\left(1 - \frac{\sigma_{ap}}{\mu_{ap}}\right) \tag{5}$$

(3)是否为单例评论人

单例评论指的是评论人仅对某个商家发表评论的行为. 正常情况下,用户会对不同商家发表评论. 而虚假评论人账号经常出现单例评论的情形. 该特征记为 SR_a:当用户为单例评论人时计 1,否则计 0.

(4)早期评论倾向

在商家注册的初期,迫切需要增加评论以吸引客户,主观上有收买水军进行评论的动机. 另一方面,初期由于评论数目较少,每条评论的评分对于商户整体评分的影响比较大. 从客观上看,在此时进行虚假评论的效率较高. 因此采用早期评论人作为模型的一个特征,用 ETF_a 表示如下:

$$ETF_a = \max_{p \in P_a} ETF_{ap} \tag{6}$$

其中 ETF_{ap} 指评论人 a 对于商家 p 的 ETF 值:

$$ETF_{ap} = \begin{cases} 0, & \text{if } L_{ap} - A_p > \beta \\ 1 - \dfrac{L_{ap} - A_p}{\beta}, & x \geqslant 0 \end{cases} \tag{7}$$

其中,L_{ap} 表示评论人 a 对商家 p 发表的最后一条评论的时间戳,A_p 指该商家第一条评论被发表,或者该商家注册的时间戳. 该函数为分段函数,其中 β 作为阈值,本文设定为六个月. 当评论人 a 对商家 p 发表的最后一条评论的时间与商家商家注册的时间差距超过 β,则认为该用户非早期评论人,该特征计为 0;如果该差距在 β 以内,则不直接标记为 1,而是对其进行量化并归一化.

(5)评论人逻辑回归类标签

为构建逻辑回归模型,需要对训练集数据进行类标记. 本文对于评论人可信的训练集数据标记 y_a 做如下定义:如果用户曾经发表过被标记为虚假评论的评论,则认为其类标签为 0,否则为 1. 需要注意的是这并不意味着该用户(评论人)发表的所有评论都是虚假评论,对其进行此种标记的目的是为了构建逻辑回归模型,并使对用户计算可信度成为可能.

3.3.2　商家可信度模型

商家可信度模型同样选用的逻辑回归算法,商家的特征有如下考虑:

(1)单例评论比例

单例评论同样在评价商家可信度中进行使用,该特征记作 SRr_p,即 p 商家所有评论中由单例评论人发表的评论数目,与所有评论数目的比例.

(2)爆发评论比例

商家评论的发表频率有时不是平滑而均匀的. 在正常情况下,特殊时间点上,商家的评论数目会比平时多,比如节假日、促销或者团购活动发生时,这时会形成评论数量高峰. 这一高峰同样会出现在商家雇佣虚假评论人提高自己商誉或者贬低对手商誉时. 因此,对商家评论数目随着时间的起伏进行观察并记录对构建商家可信度模型具有一定帮助. 本文将该指标记作 $BSTr_p$,即 p 商家所有评论中,在高峰期发表的评论数目与所有评论数目的比例.

为了探测并识别发表评论的高峰,本文采用核密度估计(kernel density estimation)[13,14] 方法进行检测. 令 $(x_1, x_2, \cdots, x_n)$ 为一连串符合密度函数 f 的独立同分布的变量值,则 f 表示为:

$$\hat{f}_h = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{8}$$

其中 $K(\cdot)$ 即"核",即一个对称但不一定为正的函数,其所有值相加为 1,$h > 0$ 是称为带宽的平滑参数. $K(\cdot)$ 为按比例变换后的核,即 $K_h(x) = \frac{1}{h}K(x/h)$. 常用的核有均匀核函数,三角核函数,伽马核函数,正太核函数等.

通过核密度估计,得到商家 p 评论发表时间的一条密度函数曲线,本文采用 Kleinberg 高峰检测模型[15],

在估计核密度的基础上还得到了密度高峰的层级值,本文采用第二层级内时间发表的评论作为密度高峰评论,将其标记为 Burst Reviews.

(3)密度差异系数

商家 p 发表评论的密度曲线中,除了高峰处之外,曲线本身起伏的程度也将继承到商家可信度模型当中. 本小节中,以 DCV_p 记改商家密度曲线的差异系数:

$$\mathrm{DCV}_p = \frac{\sigma(\text{density curve})}{\mu(\text{density curve})} \tag{9}$$

(4)评分差异系数

一个理性的消费者可以判断商品或服务的质量,并给出相应的合理评分,因此一个正常商家的不同评论评分之间的差异度应当不会非常大. 而虚假评论人出于特殊目的其发表的评论与其他评论的差异会比较大,基于此,以 RCV_p 记商家 p 所有评论评分的差异系数:

$$\mathrm{RCV}_p = \frac{\sigma(\text{scores})}{\mu(\text{scores})} \tag{10}$$

(5)平均评分与平均专项评分

平均评分 avgstar 即该商家 p 的所有评论的平均评分. 平均专项评分 avgscore1,avgscore2,avgscroe3 是餐饮业商家特有的专项评分,即口味、环境、服务的平均分数.

(6)商家逻辑回归类标签

为了构建逻辑回归模型,需要对训练集数据进行类标记. 本文对于商家可信的训练集数据标记 y_p 做如下定义:如果对该商家曾经发表过被标记为虚假评论的评论,则认为其类标签为 0,否则为 1. 同样,这并不意味着对该商家发表的所有评论都是虚假评论.

3.4 虚假评论识别模型

除了与一条评论相关的评论人可信度 P_a 与商家可信度 P_p 外,评论还具有一些其他特征:

(1)评论长度与词性分布:文本预处理阶段,计算了各种词性与标点符号在评论中所占的比例. 经实验对比,对判断评论可信度有效的词性比例特征是动词与标点符号所占的比例;

(2)评论评分与整体评分差异:将评论 r 的评分与商户整体评分之间的差异,记作 stard:

$$\mathrm{stard}_r = \frac{|r_{rp} - \overline{r_p}|}{40} \tag{11}$$

(3)评论专项评分之间的差异:对于商家的评论 r,本文计算其口味、服务、环境三个专项评分的标准差,记作 scored,作为特征. 本文认为一个理性的消费者应当会对餐饮业商家的不同方面做出理性判断,从而不倾向于给出三项专项评分一致的判断.

$$\mathrm{scored}_r = \sigma(\mathrm{score1}, \mathrm{score2}, \mathrm{score3}) \tag{12}$$

使用上述全部特征构建逻辑回归模型,计算 p-value 可知,重要的特征有相关用户可信度 P_a、商家可信度 P_p、文本长度 len_r、评论评分与整体评分差异 stard_r,评论专项评分之间的差异 scored_r 以及标点 s_w、动词 s_v 占所有词的比例,以这些词语重新构建模型并测试. 该模型可表示为:

$$P_r = h_\theta(x_r) = g(\theta_r^t x_r) = \frac{1}{1 + \mathrm{e}^{-\theta_r^t x_r}} \tag{13}$$

$$\theta_r^t x_r = \theta_{1r} x_{1r} + \theta_{2r} x_{2r} + \cdots + \theta_{nr} x_{nr} \tag{14}$$

4 虚假评论识别模型的测试

4.1 测试指标

二元分类问题,常常使用混淆矩阵来表示分类结果,矩阵的每一列代表一个类的实例预测,而每一行表示一个实际的类的实例. 一个典型的混淆矩阵如表 1.

表 1 混淆矩阵

		真实值		总数
		p	n	
预测	p'	真阳性(TP)	伪阳性(FP)	P'
	n'	伪阴性(FN)	真阴性(TN)	N'
总数		P	N	

其中,把伪阳性(FP)作为第 1 型错误(Type I Error),将伪阴性(FN)作为第 2 型错误(Type II Error). 使用准确度(Accuracy)作为评价指标:

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{TF}} \tag{15}$$

但仅用 Accuracy 在类别分布不平衡时难以真实评价模型的好坏,无法区分出第 1 型错误与第 2 型错误. 因此考虑使用敏感性(Sensitivity)和特异性(Specificity):

$$\mathrm{Sensitivity} = \frac{\text{True positive}}{\text{Condition positive}} \tag{16}$$

敏感性即真阳性率 TPR(True Positive Rate),该值越高,第 2 型错误越少,即被误分入虚假点评的真实点评比例越低.

$$\mathrm{Specificity} = \frac{\text{True negative}}{\text{Condition negative}} \tag{17}$$

特异性即 1-假阳性率 FPR(False Positive Rate),该值越高,第 1 型错误越少,即被错误分入正常点评的虚假点评比例越低. 由此,采用敏感性与特异性作为验证模型要求的评测指标. 然而由于对分类器取不同阈值时,可以得到不同的分类结果及分类器评价指标,采用 ROC(Receiver operating characteristic)曲线[16]与 AUC(Area Under the Curve)[16]值评价模型. ROC 曲线描述的是分类混淆矩阵中 FPR-TPR 两个量之间的相对变化情况. 如二元分类器输出的是对正样本的一个分类概率值,当取不同阈值时会得到不同的混淆矩阵,对应于 ROC 曲线上的一个点. ROC 曲线反映了 FPR 与 TPR 之

间权衡的情况,TPR 增长得越快,曲线越往上屈,AUC 就越大,模型的分类性能就越好. AUC 值指的是 ROC 曲线下部的面积的和,如图 2 所示,在模型达到最佳分类情形是,ROC 曲线紧贴模型的左侧和上侧,AUC 值达到最大为 1;当 AUC 在 0.5 ~ 1 之间时,模型优于随机预测;当 AUC 为 0.5 时,模型为随机预测,没有价值;当 AUC 小于 0.5 时,模型预测结果劣于随机值.

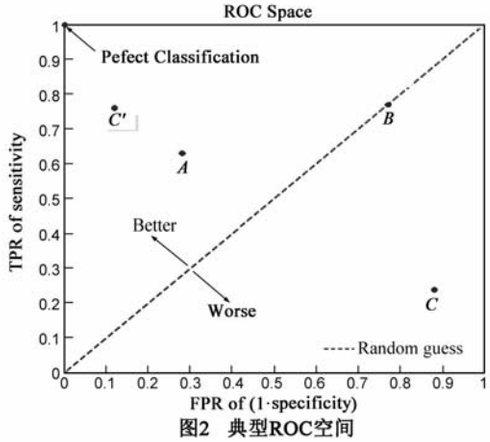

图2　典型ROC空间

综上,本文将主要使用 AUC 值作为评价指标,AUC 值越大,则模型表现越好.

4.2　点评人、商家可信度模型测试

为分别验证点评人、商家的可信度模型,将计算特征后得到的 238186 个点评人记录、109376 条商家记录分别按 7∶3 分为训练集和测试集,构建模型后进行测试,并计算结果的 ROC 曲线与 AUC 值.

对于点评人可信度模型的训练、测试数据格式如下:〈点评评分与商家整体点评评分间差异〉〈对同一商家点评评分相似性〉〈是否为单例点评人〉〈早期点评人指标〉〈点评人是否发表过虚假点评〉

即:　$x_a = (\mathrm{RD}_a, C_a, \mathrm{SR}_a, \mathrm{ETF}_a)\ h_\theta(x_a) = y_a$

对于商家可信度模型的训练、测试数据格式如下:〈单例点评比例〉〈爆发点评比例〉〈密度差异系数〉〈评分差异系数〉〈平均评分与平均专项评分〉〈商家名下是否有虚假点评〉

即:　$x_a = (\mathrm{SR}r_p, \mathrm{BST}r_p, \mathrm{DCV}_p, \mathrm{RCV}_p, \mathrm{avgscore1}, \mathrm{avgscore2}, \mathrm{avgscore3})$

$h_\theta(x_p) = y_p$

点评人、商家可信度模型的 ROC 曲线如图 3、4 所示.

用户可信度模型 AUC 值为 0.858922,而商家可信度模型的 AUC 值为 0.8792504. 可见,当把用户及商家模型看做一个分类问题时,具有较好的分类结果与精度,因此用户可信度模型与商家可信度模型均表现优良.

4.3　虚假点评识别模型测试

将 331,415 条数据按 7∶3 的数目比例分为训练集

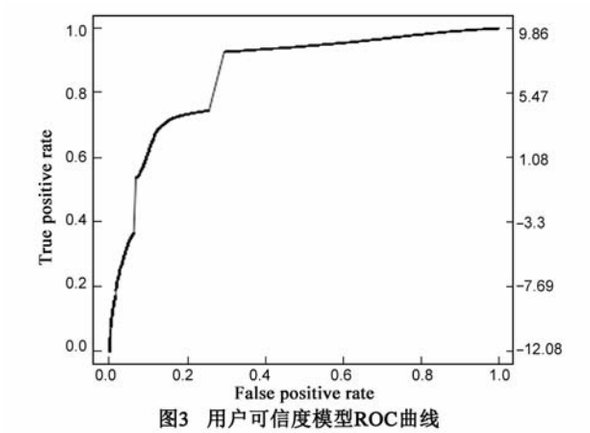

图3　用户可信度模型ROC曲线

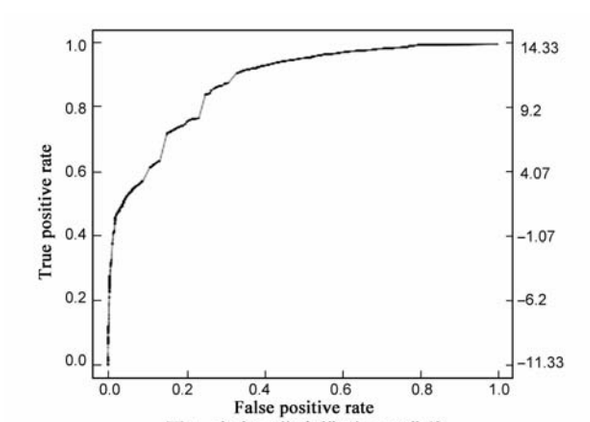

图4　商家可信度模型ROC曲线

与测试集. 对于虚假点评识别模型的训练、测试数据格式如下:

〈点评人可信度〉〈商家可信度〉〈点评长度〉〈点评评分与整体评分差异〉〈点评专项评分之间的差异〉〈词性分布〉〈点评是否虚假〉

即:　$x_r = (p_a, p_p, \mathrm{len}_r, \mathrm{stard}_r, \mathrm{scored}_r, \mathrm{s}_w, \mathrm{s}_v)$

$h_\theta(x_r) = y_r$

对虚假评论识别模型进行测试,该模型的 ROC 曲线如图 5 所示.

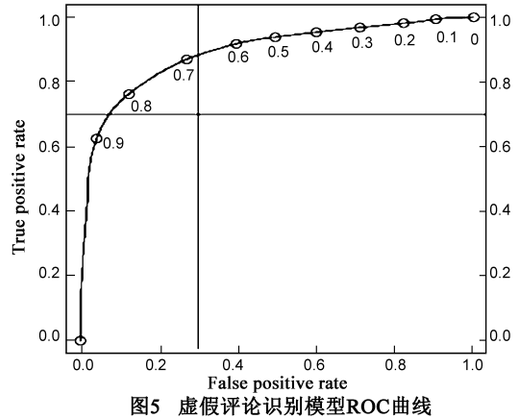

图5　虚假评论识别模型ROC曲线

计算得到 AUC 值为 0.8930167. 可见,考虑了用户可信度,商家可信度及评论文本特征的虚假评论识别模型具有较好的分类结果与精度,因此该模型表现优良.

5 总结

本文基于某网站上的评论进行研究,分析并总结虚假评论的特征,结合特征提出用户、商家可信度模型,利用评论人的行为特征、商家的特征和评论文本的特征构建了虚假评论识别模型. 使用 AUC 值作为主要评测指标,经测试模型达到了良好的区分效果.

参考文献

[1] 刘逶迤,逯万辉,丁晟春. 商品评论信息可信度研究[J]. 情报科学,2012,30(10):1556 - 1565.
Liu Weiyi, Lu Wanhui, Ding Shengchun. Research on the credibility of commodity reviews information[J]. Information Science,2012,30(10):1556 - 1565. (in Chinese)

[2] Kusumasondjaja S, Shanka T, Marchegiani C. Credibility of online reviews and initial trust: the roles of reviewer's identity and review valence[J]. Journal of Vacation Marketing,2012,18(3):185 - 195.

[3] Pan L Y, Chiou J S. How much can you trust online information? cues for perceived trustworthiness of consumer-generated online information [J]. Journal of Interactive Marketing,2011,25(2):67 - 74.

[4] Qiu L Y, Pang J, Lim K H. Effects of conflicting aggregated rating on ewom review credibility and diagnosticity: the moderating role of review valence [J]. Decision Support Systems,2012,54(1):631 - 643.

[5] Cheung M Y, Luo C, Sia C L, et al. How do people evaluate electronic word-of-mouth? informational and normative based determinants of perceived credibility of online consumer recommendations in china [A]. 11th Pacific Asia Conference on Information Systems [C]. Auckland: Bepress,2007. 69 - 73.

[6] Cheung M Y. Do People Believe Electronic Word-of-Mouth?: A Study on Factors Affecting Readers' Perceived Credibility of Online Consumer Reviews[D]. Hongkong: City University of Hong Kong,2006.

[7] Cheung M Y, Cindy M Y, Sia C L, et al. Is this review believable? a study of factors affecting the credibility of online consumer reviews from an elm perspective[J]. Journal of the Association for Information Systems,2012,13(8): 618 - 635.

[8] Ott M, Choi Y J, Cardie C, et al. Finding deceptive opinion spam by any stretch of the imagination[A]. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies [C]. Stroudsburg, PA, USA: Association for Computational Linguistics,2011. 309 - 319.

[9] Li F T, Huang M, Yang Y, et al. Learning to identify review spam [A]. 22nd International Joint Conference on Artificial Intelligence[C]. Barcelona: AAAI Press,2011. 2488 - 2493.

[10] Mukherjee A, Venkataraman V. What yelp fake review filter might be doing? [A]. Proceedings of the 7th International Conference on Weblogs and Social Media[C]. Palo Alto: AAAI Press,2013. 409 - 418.

[11] Zaki M J. Scalable algorithms for association mining[J]. IEEE Transactions on Knowledgeand DataEngineering. 2000,12(3):372 - 390.

[12] Hosmer D W, Lemeshow S. Applied Logistic Regression [M]. New York: John Wiley & Sons,2004.

[13] Rosenblatt M. Remarks on some nonparametric estimates of a density function[A]. Selected Works in Probability and Statistics[C]. New York: Springer New York,2011. 95 - 100.

[14] Parzen E. On estimation of a probability density function and mode[J]. Annals of Mathematical Statistics,1962,33(3):1065 - 1076.

[15] J Kleinberg. Bursty and hierarchical structure in streams [A]. 8th ACM SIGKDD Intl Conf on Knowledge Discovery and Data Mining,2002[C]. Edmonton, Alberta, Canada: ACM Press,2002. 91 - 101.

[16] Swets J A. Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers[M]. New Jersey: Lawrence Erlbaum Associates, Inc,1996.

作者简介

李 晶 女,1981 年生于山东威海. 北京邮电大学软件学院讲师. 研究方向为智能信息处理、数据挖掘.
E-mail: lijingjing@ bupt. edu. cn