目 录

[中文摘要： 1](#_Toc421645632)

[英文摘要： 2](#_Toc421645633)

[1 绪 论 3](#_Toc421645634)

[1.1 研究背景与意义 3](#_Toc421645635)

[1.2 主要研究现状 4](#_Toc421645636)

[1.2.1 在线评论的相关研究 4](#_Toc421645637)

[1.2.2 半监督学习的相关研究 5](#_Toc421645638)

[1.3 主要研究内容 5](#_Toc421645639)

[1.4 研究方法与技术路线 6](#_Toc421645640)

[1.4.1 研究方法 6](#_Toc421645641)

[1.4.2 技术路线 6](#_Toc421645642)

[1.5 本文结构 7](#_Toc421645643)

[2 在线评论有用性分析理论研究 9](#_Toc421645644)

[2.1在线评论 9](#_Toc421645645)

[2.1.1 在线评论概述 9](#_Toc421645646)

[2.1.2 在线评论的特征 10](#_Toc421645647)

[2.2 文本挖掘 11](#_Toc421645648)

[2.2.1 文本挖掘的概述 11](#_Toc421645649)

[2.2.2 文本挖掘过程 12](#_Toc421645650)

[2.2.3 文本分类 12](#_Toc421645651)

[2.3 半监督学习 14](#_Toc421645652)

[2.3.1 自我训练算法 15](#_Toc421645653)

[2.3.2 协同训练算法 16](#_Toc421645654)

[3 基于半监督学习的在线评论有用性分析研究 18](#_Toc421645655)

[3.1 基于半监督学习的在线评论有用性分析整体流程 18](#_Toc421645656)

[3.2 在线评论数据获取 18](#_Toc421645657)

[3.3 在线评论数据预处理 19](#_Toc421645658)

[3.4 在线评论的特征选择 20](#_Toc421645659)

[3.5 在线评论数据分类 21](#_Toc421645660)

[3.5.1 基于取样的非均衡数据分类方法 22](#_Toc421645661)

[3.5.2 基于集成学习的数据分类方法 22](#_Toc421645662)

[3.5.3 半监督学习与非均衡分类结合 24](#_Toc421645663)

[3.6 结果评价指标 25](#_Toc421645664)

[4 实验设计与结果分析 27](#_Toc421645665)

[4.1 实验设计 27](#_Toc421645666)

[4.1.1 实验环境 27](#_Toc421645667)

[4.1.2 实验数据 27](#_Toc421645668)

[4.1.3 实验流程 29](#_Toc421645669)

[4.2 实验结果分析 30](#_Toc421645670)

[4.2.1 实验结果整体分析 30](#_Toc421645671)

[4.2.2 基于结构特征的实验结果统计分析 33](#_Toc421645672)

[4.2.3 基于混合策略的实验结果统计分析 37](#_Toc421645673)

[4.2.4 结构特征下不同添加样本对实验结果统计分析 41](#_Toc421645674)

[4.2.5 混合策略下不同添加样本对实验结果统计分析 46](#_Toc421645675)

[4.2.6 混合策略和结构特征的对比分析 51](#_Toc421645676)

[结 论 53](#_Toc421645677)

[谢 辞 54](#_Toc421645678)

[参考文献 55](#_Toc421645679)

电子商务中基于半监督学习的在线评论

有用性分析

摘 要：互联网技术的快速发展，推动了电子商务的普及，消费者越来越倾向将于将日常活动迁移到网络上，在线评论已经成为消费者了解网上商品信息的一个重要途径，然而在线评论评论数量巨大，而且更新速度快，造成了消费者信息过载，良莠不齐的评论质量需要消费者花费大量的时间去筛选有用的评论，目前消费者迫切需要一种方法对在线评论的有用性进行分析。传统的分析方法需要标记大量的样本，提高了实验的成本，而未标记的样本中存在着大量隐藏的信息未被利用。因此，本文提出在电子商务中基于半监督学习方法方式利用少量标记样本和大量没标记的信息对在线评论进行有用性分析。首先，本文以豆瓣电影用户发表的长评论作为数据源，其次，对数据集去除空值、筛选、分词和去停用词等预处理操作；然后，以半监督学习中协同训练和自我训练方法与非均衡数据分类方法相结合的方式对在线评论的有用性进行了实验对比分析，实验对比的结果验证了本文提出的方法的可行性和有效性。

关键词：在线评论；有用性分析；半监督学习；非均衡数据分类；电子商务

Research on Online Review Helpfulness Based on Semi-supervised Learning in E-commerce

Abstract: The rapid development of Internet technology promotes the popularity of e-commerce so that consumers are increasingly inclined to online shopping. Consumers primarily get goods information through online reviews. However,the huge amount of online reviews with fast update rate results in a information overload to consumers. Therefore, it is necessary for a method to analyze the usefulness of online reviews.Traditional methods marking large sample not only raises the cost, but also causes that a lot of information hidden in the unmarked sample is not used.Therefore, we propose a method based on semi-supervised learning in e-commerce that uses little marked samples and plenty of information hidden in the unmarked sample to analyze the usefulness of online reviews.Firstly, we use the reviews on movies in douban as a data source. secondly, we preprocess the data collection ,such as removing the null value, filtering and segmentation. Then, combining the learning collaborative training and self-training methods in a semi-supervised with unbalanced data classification to analyze and compare the usefulness of online reviews. The results shows that the method we propose is feasible and effective.

Keywords: Online Reviews; Analysis on Helpfulness; Semi-supervised Learning; Unbalanced Data Classification; E-commerce

# 1 绪 论

## 1.1 研究背景与意义

根据中国互联网信息中心第35次全国互联网发展状况调查统计报告所示，截至2014年12月，中国网民规模达6.49亿，全年共计新增网民3117万人。互联网普及率为47.9%，较 2013年底提升了2.1个百分点[1]。随着互联网的快速发展和普及，越来越多的用户使用互联网提供的服务，其中社交媒体发展迅速。用户往往利用微信、QQ、论坛、微博、评论网站等社交媒体平台发表自己对于热点事件、商品的观点或参与讨论，这些在线评论观点包含着用户针对产品或者服务的用户体验和态度。通过浏览在线评论信息，用户可以了解已购买商品的用户对商品的评论，以此作为用户购买决策时的参考依据，企业可以根据在线评论指导企业的生产经营活动。但是在线评论的数量庞大，用户往往面临上百甚至更多的评论，存在信息过载的问题，并且评论质量的良莠不齐，还可能受到商业雇佣的评论员发表不切实际的评论干扰用户判断。

传统的对在线评论的分析方法可以分为文本情感挖掘和机器学习。文本情感挖掘需要深度标注文本的词性、句法、语义等信息，并且目前的语料库还没达到比较好的规模。在机器学习领域中，传统的学习方法有两种：有监督学习和无监督学习。有监督的学习方法需要大量标注样本，忽略了未标记样本信息隐藏的大量待发掘的信息，而无监督学习方法精度不高。半监督学习是监督学习和无监督学习相结合的一种学习方法。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。半监督学习对于减少标注样本代价，提高学习性能具有非常重大的实际意义。目前，根据半监督分类算法的实现方法现有的典型方法大致分为五种：基于生成模型的半监督分类方法、自我训练方法、协同训练方法、基于图的半监督分类方法、直推式支持向量机方法。本文使用自我训练和协同训练对样本分析。协同训练算法利用两个训练器对文本进行训练。避免了自我训练迭代过程中将错误越累积越大，提高了可靠和分类精度。由于传统的数据分类方法假定在各个类别数量总体上相等的情况下，本文将原始的数据分类方法分为有用、没用后，两类之间的数据存在着显著的不均衡，传统的数据处理方法并不适用在线评论数据。因此本文将基于抽样的和集成学习的分类方法引入到在线评论分析中。

综上所述，本文提出在电子商务中基于半监督学习与非均衡数据分类方法相结合的方式对在线评论进行有用性分析。首先，本文利用Java编写的数据抓取程序从豆瓣电影上抓取用户发表的对电影的长评论作为实验的数据源；其次，对原始在线评论数据进行了去除空评论内容、类别筛选、分词、去停用词等预处理操作，并且选择了100个中文文本特征并编程实现，然后，使用基于半监督学习与非均衡的数据分类方法相结合的方式，以5次2倍交叉实验方法对本文提出的方法进行实验分析，最后对实验结果进行了分析， 实验结果验证了在电子商务中基于半监督学习方法对在线评论做有用性分析是可行与有效的。本文研究中将半监督学习方法引入到在线评论分析，因为在线评论呈现出不均衡的特点，所以将半监督学习和非均衡数据分类相结合的方式进行评论的有用性分析，拓展了对在线评论的分析方法。

## 1.2 主要研究现状

本文主要涉及在线评论的分析和使用半监督的学习方法等相关内容，相关内容的国内外研究现状具体如下：

### 1.2.1 在线评论的相关研究

在2001年，Chatterjee首次提出在线评论这一概念，在线商品评论对于营销活动和企业管理的价值便被越来越多的营销领域学者关注[2]。相比较国外的研究，国内对在线评论的研究起步较晚。国外相关研究学者指出在线评论已经成为消费者在网络中进行消费做决策时的重要参考依据，同时在线评论可以将消费的反馈信息传递给企业，企业可以利用这些信息对日常活动进行调整、指导。国内许多学者指出，网上在线评论不仅可以以相对较低的成本为公司收集到相关信息，也可以有效传播企业信息。这是建立用户和企业间良好信任的方式。消费者可以通过阅读在线评论了解到企业的信誉、产品质量、服务和使用产品的体验。目前基于在线评论的研究还是一个新兴领域，主要涉及评论主题的识别、评论发布者情感分析和结果可视化、评论中产品特征的提取等。

综上所述，对于在线评论的影响，国内外学者对其的研究涉及许多因素，已形成相对来说较完整的理论体系。但是目前关于在线评论对电子商务的影响的研究相对其他方面相对较少。

### 1.2.2 半监督学习的相关研究

相关研究者通常将1994年Shahshahani、Landgrebe[3]的研究工作视为对半监督学习研究的开始标志，事实上在20世纪80年代，有一些研究人员就已经意识到无标记样本的价值，因为当时相对成熟的机器学习相对较少的利用无标记样本的价值所以对半监督学习的研究还比较少。然而随着技术的不断发展，以及人们对无标记样本或少量的标记样本的利用的需求越来越强烈，研究者们对半监督学习的研究逐渐增多，而半监督也成为新的研究方向。在当前对半监督学习的研究正从广度、深度两个方面不断的深入，一方面，研究者们在半监督学习中引入新的算法，提高特征提取和分类的性能；另一方面，研究者们不断在新提出的监督学习算法中进行扩展，用以提高在半监督学习算法的性能。当前的半监督学习通过与其他技术相结合的方式，已经不单单依靠数据训练来学习了。

总之，半监督学习已经成为当前研究的热点领域之一，在期刊中关于半监督学习的文章也越来越多，关于半监督学习的专题会议也呈现上升的趋势。在模式识别、机器学习等学科半监督学习对于推动其发展也起着积极的作用。

## 1.3 主要研究内容

本文研究的主要内容是对电子商务中的在线评论进行有用性分析，由于电子商务中可以比较容易的获取大量的无标记的样本以及以较低的成本对样本进行少量的标记，所以本文首先提出基于半监督学习的方法对在线评论进行分析，其次，本文所抓取的在线评论呈现数据不均衡的特点，本文将分均衡数据分类方法引入到在线评论的分析中。使用Java编写程序从豆瓣电影抓取用户对于电影的长评论作为数据源，再对文本进行预处理，通过实验验证在电子商务中基于半监督学习的在线评论有用性分析的有效性。

首先，通过网络和图书馆查阅在线评论分析、文本分类和半监督学习相关的资料，了解电子商务中在线评论的特点、熟悉半监督学习中自我训练和协同训练的流程和算法，学习了非均衡数据的分类方法，以及了解了在线评论和半监督学习研究现状；

其次，编程实现从豆瓣电影网站抓取用户关于电影的长评论，为在线评论分析做准备工作，将抓取的在线评论集划分为有用和没用两类，使用编写程序实现了选取的100个中文特征，并完成所有特征方法的单元测试；

再次，根据中科院的分词系统对在线评论数据集进行中文分词和去除连词、语气词、标点符号等对分类结果无用的停用词；执行编写的文本特征程序生成了ARFF文件，编程实现调用Weka中的几种分类方法对以上语料库进行分类；

最后，对在线评论进行了试验，对生成的结果数据进行了处理，通过图表对几种分类方法进行对比分析，以AverageAccuracy和AUC两种评价指标对最后的分类结果进行分析比较，并对实验结果进行说明。

## 1.4 研究方法与技术路线

### 1.4.1 研究方法

本文主要使用了文献研究法、实验法、理论分析法、比较分析法、等方法相结合来对电子商务中的在线评论的有用性分析。首先，通过互联网工具在知网、万方等数据库查阅了在线评论、文本挖掘、文本分类、半监督学习、非均衡分类等相关资料和文献，了解在线评论分析、文本分类等领域的相关问题，学习自我训练、协同训练算法以及分均衡数据分类的理论方法，从而掌握自己所要研究的课题。本文实验以两种策略即仅使用结构特征与混合策略相结合的方式，对半监督学习与非均衡数据分类相结合的方式与基本分类器方法进行对比，并分析了相结合的方法相对与基本分类器的效果提高百分比，分析了在不同比例的标记样本对各个方法实验结果的影响，实验中也考虑了不同添加样本对实验结果的影响，实验结果均以两种指标作为评价参数。

### 1.4.2 技术路线

本文在研究过程中主要采用了图1.1所示的技术线路图，首先通过文献分析和资料收集了解相关理论基础，再使用Java编写的程序从豆瓣电影上抓取用户评论，得到数据集；然后对原始的数据集去除空值、并分为有用和没用两类别、分词、去除停用词以及格式转换等一系列预处理；再使用Weka软件进行分类，得到分类结果；最后根据AA、AUC指标对分类结果进行分析。

图1.1 研究技术线路图

## 1.5 本文结构

本文总共分为五个章节，主要的内容如下所示：

第一章 主要介绍了本文的研究背景与意义、国内外对于在线评论和半监督学习的相关研究、主要的研究内容，最后介绍了研究所用的方法与技术路线以及本文的结构。

第二章 主要介绍在线评论、文本挖掘和对半监督学习的简述、以及自我训练和协同训练算法介绍分析；

第三章 主要介绍基于监督学习的在线评论有用性分析的研究的框架、数据获取模块、数据预处理模块、非均衡的数据分类模块、半监督学习与分均衡的数据分类方法的结合、结果分析模块。

第四章 主要由实验设计与结果分析构成。实验设计主要介绍了实验环境、如何解决在线评论的数据源以及实验流程；然后对实验的结果进行了整体分析、不同分类器有用性的对比分析。

最后 对本文在线评论的分析作了总结，并指出了文章中存在的不足。

# 2 在线评论有用性分析理论研究

## 2.1在线评论

随着Web2.0的发展和普及，越来越多的网站开始引入在线评论模块，在互联网中的各个方面都积累了大量的评论数据，如何从这些文本中发掘有用的信息已经成为各个领域的共同需求。本节对在线评论做了概述，介绍了在线评论的特征。

### 2.1.1 在线评论概述

随着电子商务的兴起，在线评论的价值不断提高，也越来越多的受到研究者的关注。在线评论(Online Reviews)是指消费者将对商品质量、性能、使用体验等的评价信息以文本的形式发布在网络上，这些评价包括对产品的赞美、抱怨，或个人对特定产品或服务的购买和使用感受[4]。根据2013年针对网购用户购买不熟悉的商品时主要考虑因素的统计调查显示，在所有考虑到的因素当中，在线用户评论以32.90%的投票比例排在首位，它表明消费者在网络消费时在线评论已经成为其购买决策的重要参考依据[5]。在线评论信息一般包括发布在线评论的用户信息、对评论评价、评论的内容和评论发表的时间等组成，在线评论的发布者可以采取以匿名的形式发表评论，网站还可以显示评论发布者的等级，通常评论发布者的等级越高，他所发表的在线论的可信度也就越高。用户对商品的使用体验可以以评论等级的形式直观的反映给评论的阅读者。评论的等级一般可以分为好、较好、中、较差、差；评论内容是评论者对商品的质量、使用体验的描述；评论发布时电子商务网站自动记录评论的发布时间。

在电子商务中，在线评论可以降低消费者和卖家双方之间的信息不对称，这种积极有效的反馈机制还可以增加消费者对卖家的信任，降低消费者对风险的感知，在线评论在为买卖双方带来便利的同时也面临越来越严峻的挑战。 首先，因为在电子商务中网络平台允许用户以匿名的方式发布在线评论信息，所以有部分用户会利用这一点发布不真实的评论，会影响用户通过阅读在线评论对产品的判断。其次，在线评论价值不断提高，促使一些以发表在线评论的人员产生，他们被称作五毛党。由于受到商业公司、网站等机构的雇佣，在各种评论发布渠道上发表受雇佣者指导的各种评论，为了避免被发现，他们通常使用普通网民的身份发帖。大量的造假现象使得商品评论信息的可信度无法得到保证，这些可信度的评论误导了消费者，扰乱了电子商务市场。最后，消费者产生的在线评论的数量越来越大，以至于消费者不可能通过阅读所有内容了解其中有价值的评论，这将耗费其大量的时间成本与精力，也使其越来越难以做出对自己利益最大化的购买决策。筛选出有用的在线评论可以帮助消费者克服在线购买所承担的风险，也有利于消费者之间分享个人实际使用产品的体验，将有价值的在线评论内容给予其他消费者的借鉴。因此，为了帮助消费者识别有用的在线评论，提高用户购买体验，电子商务网站应该识别出对消费者有有价值的评论，使得消费者购买决策更简单。因此，对在线评论有用性分析的研究有着重要的理论意义和实践价值。

### 2.1.2 在线评论的特征

通过网络浏览在线评论以及参考其他学者的研究，本文总结了在线评论呈现出的特征即内容重复性、内容简短单句话较多、非结构化、匿名性、动态性、以较低的成本获取少量的评论信息[6]。

内容重复性指的是消费者购买商品后，会带有强烈的个人情感倾向，为了在线评论中表达自己强烈的情感，往往会在评论商品的时候，不断的重复某个词语或者某句话；内容简短单句话较多是指消费者在网上发表购买评论往往评论的内容比较短，大多数评论只有区区几十个字，字数多一点的评论也不过百余字，并且这种评论通常没有主语；内容结构不规范是指由于大多数消费者购买商品后，发表评论时会使用日常中的口语，俚语，同音字，缩写等将自己体验商品之后的真实感受表达出来。消费者在书写商品评论时输入汉字时常识上的拼写错误，以及在线评论的表达自由，社会敏感词汇的限制，评论者通常会输入一些噪声、非规范词、非规范符号和规范语言格式都会导致这种内容的不规范；匿名性是指参与评论的消费者可以采取匿名的方式发布评论。对于评论者来说，采用匿名的方式可以避免所发表的负面评论给自己带来负面的影响，所以采用匿名的评论者可以提供对商品，服务真实的意见和想法。当然采用匿名的的不利在竞争对手可以自己或者雇人发表虚假信息，这将会误导其它的消费者阅读商品评论做出自己的决策；动态性是指商品的评论者可以在发表评论之后追加，修改评论信息；可以以较低的成本获取少量的评论信息是指在线评论的数量巨大，而且产生的速度较快，抓取所有的评论数量不仅耗费巨大的人力、物力、财力，但却可以以较低的成本获取少量的评论信息。

## 2.2 文本挖掘

随着互联网的迅速发展和深入应用，互联网中产生了大量非结构化的文本，传统的数据挖掘技术并不能很好的处理文本，挖掘潜在的信息；文本挖掘的对象是非结构化的文本对象，能够较好的对文本进行分析。本小节对文本挖掘过程、文本分类作了简介。

### 2.2.1 文本挖掘的概述

在电子商务日常活动中会产生数量巨大的有用没用的信息，在这些信息中存在着大量没有被发掘的信息，并且这些文本内容是非结构化的数据，目前，从这些海量的数据中发掘隐藏的有用信息已经吸引了许多研究者的注意。文本挖掘(Text Mining, TM)是以计算语言学、统计数理分析为理论基础，结合机器学习和信息检索技术，从文本数据中发现和提取独立于用户信息需求的文档集中的隐含知识，它是一个从文本信息描述到选取提取模式，最终形成用户可理解的信息知识的过程[7]。本文将使用文本挖掘的知识从从这些文本中获取对用户有价值的信息。

文本挖掘主要有以下4个特点[7]：（1）文本挖掘需要处理成千上万的文档集合；（2）文本挖掘算法需要具有健壮的鲁棒性，因为文本内容结构是非结构化的，而且含有大量的噪声；（3）文本挖掘需要挖掘的知识是未知的，并且几乎全部隐藏在海量的文本数据中的；（4）由于文本数据来源于真实世界用户之间的日常交流，所以文本挖掘所得到的知识是具有现实的意义和可能潜在的价值的，而这些知识又是某些用户急切需要的。

数据挖掘(Data Mining, DM)是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中采掘出隐含的、先前未知的、对决策有潜在价值的知识和规则[8]。一般认为文本挖掘是数据挖掘领域的一个分支，文本挖掘从技术本质上可以认为是数据挖掘和信息检索两门学科结合的产物。文本挖掘与数据挖掘的区别[9]：文本挖掘的数据源通常是数量巨大的文档，并且这些文档中的数据通常是非结构化或者半结构化的，没有确定的规则，特征数量往往达到几万甚至几十万，文本挖掘需要对这些文本进行预处理、特征选择等，从中挖掘出有用的新知识。数据挖掘的目标是存储在数据库中的结构化的数据，利用表与表之间的关系和存储结构挖掘出潜在的知识。

### 2.2.2 文本挖掘过程

文本挖掘的一般过程和步骤包括：获取大量的文本文档集合、为了提高文本挖掘的效率对所得到的文本文档进行预处理如去除噪声和分词等、文本内容特征的提取、然后对文本进行分类和聚类、生成知识模式、最后将生成结果通过可视化展示出来，便于用户的理解和应用[10]。具体的过程如图2.1所示。



图2.1 文本挖掘的基本过程

### 2.2.3 文本分类

文本分类（Text Categorization, TC）作为文本挖掘技术的一项重要内容，其目的是使机器自动学习一个分类模型或者分类函数，这个分类函数能够把未知类别的文本自动地映射到众多类别中的具体一个类别中，使得查询效率更高、速度更快。文本分类是指在给定分类标准的情况下，把待分类的文本内容机器自动地确定文本所属类别的过程。将文本分类应用于在线评价的有用性分析研究已经越来越广泛了。文本分类一般包括以下几个步骤：在训练阶段有获取训练文本、文本预处理、文本特征的提取、训练分类器、输出分类结果，测试阶段和训练阶段的主要阶段是获取测试文本，并对测试文本做预处理，后面的步骤和训练阶段大致相同。具体过程如图2.2所示。



图2.2 文本分类过程

文本预处理：去除原始文本数据中的噪声数据，将语料库数据转化为统一的格式，方便后续工作；在本文中对抓取的原始数据去除空值，将评论分为有用和没用两组，分词以及去除文本中停用词等。去停用词是指从文本数据中剔除掉那些出现频率过高但本身的类别区分度较小的词[11]，或者指保留文本数据中的名词、形容词、动词和副词等，去除那些对分类没有意义的词。

文本的模型表示：使用数学模型将文本文档表示成机器可以识别的统一格式，方便后面提取特征工作和分类器的训练；

文本特征的选择：由于文本非结构化的特点，相比于数据库中的数据，文本的特征数量往往达到几万甚至更高，所以需要选择最能够表示所有文本的特征，以此来降低文本的特征维度；

训练算法：选取适当的数学算法来对已选训练样本进行机器学习，得到一个分类器训练模型，再使用这个训练模型对测试样本的文本数据进行分类；

实验结果分析和性能评价：选取适当的参数对分类器的性能以及分类结果的准确性进行比较分析。

目前，文本分类中常用的分类算法主要分为基于统计的、基于连接的和基于规则的方法。其中基于统计的方法包含：朴素贝叶斯(Nai`ve Bayesian，NB)[12]、K最近邻节点(K-Nearest Neighbor，KNN)[13]、类中心向量法[14]、支持向量机(Support Vector Machine， SVM)[15]、最大熵模型[16]等；基于连接的方法有：神经网络(Biological Neural Networks，BNN)[17]；基于规则的方法有：决策树(Decision Tree，DT)[18]和基于关联规则法[19]等。

## 2.3 半监督学习

当前信息处理处理领域不断有新的技术被引入，获取无标记的样本已经能容易的实现；获取有标记的数据需要相关领域的专家对数据进行标记，而这种对数据的标记成本昂贵而且耗费时间，如果对所有的数据进行标记成本将显著提高。在机器学习领域中，根据训练集中有无标注样本将机器学习具体细分为有监督学习、无监督学习和半监督学习。

有监督需要用到大量高质量的有标记数据来训练分类器，虽然能够达到较高的准确率，但是标记数据需要大量的时间、财力、人力、物力等，而非监督方法需要耗费大量的计算资源，最终生成的类别也可能与预期不同，所以基于如何以较低的成本，并且能够利用少量标记样本所隐藏的信息和大量的无标记样本的考虑，半监督学习这种相对折中的学习方式被提出来了。半监督学习主要考虑如何利用少量的标记样本和大量的未标记样本学习对训练器进行训练。半监督学习不仅降低了标记样本的成本，还能够有效的利用未标注样本中所包含的信息。半监督学习在文本分类、图像分类、邮件过滤、机器翻译、主题词识别、词性标注等方面取得了广泛的应用，半监督学习框架图如图2.3。目前，根据半监督分类算法的实现方法现有的典型方法大致分为五种：基于生成模型（Generative Model）的半监督分类方法、自我训练方法（Self-Training）、协同训练方法（Co-Training）、基于图的半监督分类方法、直推式支持向量机方法（Transduction SVM，TSVM）等[20]。本文采用基于自我训练方法（Self-Training）和协同训练方法（Co-Training）的方法对在线评论做有用性分析。



图2.3 半监督学习框架图

### 2.3.1 自我训练算法

在使用自我训练(Self-Training)处理分类问题时，通常将训练样本分为已标记样本和未标记样本，首先利用一部分已标记样本训练出一个分类器，然后利用已训练出的分类器对未标记样本进行分类，每一个未标记样本都有一个标记置信度，设置一个阈值将置信度较高的未标记样本，连同其标记，作为已标记样本加入分类器的训练样本集。这个过程不断重复，分类器不断的更新自己的训练样本集，更新自身的分类参数，直到达到某个停止条件。



图2.4 自我训练的实现过程

值得注意的是，在自我训练中，分类器使用自身的对未标记样本的预测功能来进行自我学习和自我训练，假设有一个来自未知分布的训练样本集S，包括己标记样本集三和未标记样本集U，其中已标记样本集为L={(xl，Y1)，(x2，y2)，…… (，)}，未标记样本集U={，，…… }，其中，均为n维向量，为已标记样本，的标记类别，，分别为L和U的大小，即他们分别所包含的样本数。具体的自我训练算法的过程如图2.5。

|  |
| --- |
| 输入：已标记训练样本集L  未标记训练样本集U  过程：从U中随机抽取u，个样本组成训练池  迭代： 利用L训练一个分类器h  使用h标记出中p个正样本和q个负样本  将这些样本加入L，更新L  从U中随机抽取(p+q)个样本来更新  停止条件：训练池中已经没有未标记样本存在 |

图2.5 自我训练算法流程

自我训练算法的优点是简单易行，可以与任意的分类器h结合使用，而不用了解分类器内部的工作机制，无论这个分类器是简单还是复杂。当然它的缺点是在其迭代过程中可能会将错误的标记样本引入，在其不断的循环过程中这种错误可能会逐渐累积起来，并使的错误被放大。

### 2.3.2 协同训练算法

协同训练是一种比较常用的半监督学习方法，它最初由是Blum等[21] 在20世纪90年代提出的。他们假定对于给定的标注样本集存在两个充分冗余的视图，即符合下列条件的两个属性集：每一个属性集都能够很好地描述该问题，也就是说如果训练集比较充足，在每个属性集上都能够训练出一个强分类器；两个属性集彼此之间条件独立。基本思路为：在两个视图上基于已标注样本构建两个分类器，然后分别对未标注样本标注，且从中选取一些置信度高的添加到已标注样本集中，将更新后的已标注集作为训练集，重新训练分类器，重复这一过程直至满足条件时结束。协同训练算法过程描述如图2.7所示。

由于本方法是利用两个学习器对同一个无标记样本进行标记，那么很明显的，如果两次标记结果一致的话，该样本被标记正确的置信度就比较高，作为新的有标记样本也就更有价值。协同训练算法要求给定的训练样本能够分为两个相互独立的特征集合，并且每个特征集合都能训练出自己的分类器，然而在实践中这两个条件很难满足。



图2.6 协同训练算法的工作流程

|  |
| --- |
| 给定样本集L和U  L为有标记样本  U为无标记样本  从U中选定 u 个无标记样本 ；  Loop for k 次循环  使用 L 训练得到分类器；  使用 L 训练得到分类器；  使用标记u个无标记样本，得到结果为；  使用标记u个无标记样本，得到结果为；  对比和， 将标记结果一致的样本放入集合；  将从U中移出添加至L中；  使用学习器在L上进行训练得到最后的结果。 |

图2.7 协同训练算法流程

# 3 基于半监督学习的在线评论有用性分析研究

文中所做的分析流程的详细步骤可以分为数据获取模块、数据预处理模块、特征选择、非均衡数据分类模块、结果分析模块等内容，相关内容的详细介绍如下。

## 3.1 基于半监督学习的在线评论有用性分析整体流程

对在线评论进行有用性分析时，主要有以下几个模块：（1）在线评论数据获取，该模块主要工作是利用Java编写程序从豆瓣电影抓取用户对电影的长评论；（2）在线评论数据预处理，该模块主要是对数据获取模块得到的数据进行分词、特征提取和特征选择等预处理工作；（3）在线评论数据分类，该模块主要是调用分类方法对预处理后的数据集进行分类；（4）在线评论结果分析，该模块主要是利用评价指标对分类结果分析比较；具体流程如图3.1所示。

## 3.2 在线评论数据获取

该部分主要是使用Java编写程序抓取数据，并且将抓取到的长评论，好评数量，差评数量等关键数据存入到数据库中。根据豆瓣电影评论URL页面呈现一定的规律：前半部分为固定的字段(<http://movie.douban.com/subject/>)，后半部分也为固定的结构(/reviews)，中间部分为8到9位的数字；所以本文采取遍历10000000到100000000之间的数字，与前段和后段结果组合成新的URL，通过这个URL获取网页HTML内容，使用正则表达式匹配、获取展开评论的URL地址，然后得到此URL地址的HTML内容，在此评论的URL页面中，再次使用正则表达式匹配有用(gNum)和没用(bNum)的数量，评论的内容(content)，如果没有匹配到有用和没用的数量，则将有用和没用记为零，然后将有用和没用的数量以及抓取到的内容存入MySQL中douban数据库。

图3.1 在线评论有用性分析整体流程图

## 3.3 在线评论数据预处理

数据预处理模块主要包括以下三个工作：（1）评论筛选；以0.6作为阈值，将评论分为有用和没用，并将分类后的结果分别保存到有用和没用文件夹下，每个文件下的文本分别以数值递增的方式命名。（2）分词；分词主要是利用中国科学院计算机所编写的中文分词工具ICTCLAS进行分词。（3）去停用词；主要利用哈工大信息检索中心提供的中文停用词表去除数据集中如副词、介词、感叹词、助词、拟声词等对分类效果毫无作用的停用词。（4）特征选择。

## 3.4 在线评论的特征选择

本文从豆瓣电影中抓取用户发表的长评论作为数据源，根据Rong Zheng等[22]等人以往对文本研究和分析，本文整合四种功能融入功能集：词汇、句法、内容特有的、结构特征。其中词汇功能可进一步分为基于字符的和基于词的特点；句法功能包括功能词、标点符号和词性，在句子层面可以捕获一个作者的写作风格，句法特点的识别能力是从人组织句子的不同习惯而得；结构特征表示一个作者组织一段文本的布局方式；最后内容特别特征是为网上的消息重要的鉴别特征，这样的特征的选择取决于具体的应用领域，在网络上，一个用户可能会经常发布涉及较少在线留言而不同主题用户的范围可能分发不同主题的消息，所以特殊字或人物密切相关的特定主题可以提供一些有关作者的身份线索。本文在表3.1、表3.2、表3.3、表3.4详细介绍了基于单词的特征、基于字符的特征、基于语法的特征、基于结构的特征。

表3.1 基于单词的特征

|  |  |
| --- | --- |
| 中文单词总的数量 | Simpson’s D measure\* |
| 平均每句的单词数量 | Sichel’s S measure\* |
| Hapax legomena\* | Brunet’s W measure\* |
| Hapax dislegomena\* | Honore’s R measure\* |
| Yule’s K measure\* |  |

注：“\*”可以参考Tweedie和Baayen(1998)[23]发表的论文。

表3.2 基于字符的特征

|  |  |
| --- | --- |
| 汉字字符总的数量 | 空格字符总的数量 |
| 字母的总数量 | 制表键总的数量 |
| 大写字母的总数量 | 汉字字符总的数量 |
| 数字的总的数量 | 字母的总数量 |

表3.3 基于语法的特征

|  |  |
| --- | --- |
| 符号 ，出现的频率 | 符号 ；出现的频率 |
| 符号 。出现的频率 | 符号‘出现的频率 |
| 符号 ? 出现的频率 | 符号 ” 出现的频率 |
| 符号 ! 出现的频率 | 每个功能词语出项的频率 |
| 符号 : 出现的频率 | 中文功能词 |

注： 中文功能词参考表3.5。

表3.4 基于结构的特征

|  |  |
| --- | --- |
| 评论中所包含的行总数量 | 评论中平均每段的句子数量 |
| 评论中句子的总数量 | 评论中平均每段的字符数量 |
| 评论中段落的总数量 | 评论中平均每段的单词数量 |

表3.5 中文功能词

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 我 | 偶 | 他 | 我们 | 偶们 | 咱们 |
| 你们呢 | 他们 | 的 | 地 | 得 | 着 |
| 了 | 过 | 啊 | 呀 | 哎 | 呢 |
| 吧 | 哦 | 喔 | 呵 | 呵呵 | 也 |
| 也许 | 都 | 又 | 是 | 就 | 一个 |
| 以后 | 然后 | 然而 | 虽然 | 但是 | 到底 |
| 噢 | 随着 | 不然 | 后来 | 之后 | 总之 |
| 直到 | 往往 | 其实 | 反正 | 觉得 | 我想 |
| 认为 | 为什么 | 什么 | 怎么 | 怎样 | 难道 |
| 特别 | 却是 | 确实 | 的确 | 要不 |  |

## 3.5 在线评论数据分类

本文对从豆瓣电影抓取的电影长评论进行分类之后，有用评论的数量和没用评论的数量比例接近1:10，在线评论数据类别明显的呈现出不均衡的特点。在机器学习领域中，传统的分类方法是在假定样本数量在总体上分布是相当的，所以传统的分类方法并不适合对本文所抓取的数据进行分类[24]。针对此种情形，本文使用非均衡数据分类方法中基于取样的分类方法和基于集成学习的分类方法与半监督学习相结合的方式对实验数据进行有用性分析。

### 3.5.1 基于取样的非均衡数据分类方法

基于取样的非均衡数据分类方法主要分为欠随机取样（Under Random Sampling）[25]和过随机取样（Over Random Sampling）[26]以及合成少数类过取样方法（Synthetic Minority Over-sampling Technique），缩写为SMOTE[27]。其中US方法和OS方法是按照样本数量的影响划分的。欠随机取样方法主要是通过人为的减少多数类样本来提高少数类的分类性能，由于减少了多数类样本的数量，并且在这些多数类中包含有分类的相关信息，以致无法充分利用现存的信息；过随机取样方法主要是通过人为的增加少数类实例来提高少数类分类性能，最简单的过随机取样方法是直接复制一些少数实例。它的缺点是引入新的训练实例虽然增加了引入样本的数量，但是并没有新的信息，所以可能会导致过度拟合；其次还会增加分类器分类的时间，从而增加了训练的成本。

虽然US和OS都能够达到减小训练数据的非均衡程度，但是由于其本身所存在的不足，于是一些学者提出SOMOTE方法，它的基本思想是为了达到减小少数样本和多数样本数量之间的不均衡目的，在相距较近的少数类样本之间加入“合成”样本。SMOTE方法的不足是近邻的选择策略具有一定的盲目性[28]。SMOTE方法假设少数类实例的近邻仍然全部都是少数类，它没有考虑到样本实例的真实分布情况，如果在少数类样本附近存在一个多数类样本，那么按照SOMTE方法的假设，它会将多数类视作少数类，这显然是不合理的。

### 3.5.2 基于集成学习的数据分类方法

集成学习（Ensemble Learning）是指将多个单独的分类器以某种算法（比如加权，组合得票等）结合起来，通常当单个分类器的错误率小于0.5时，组合起来的分类器的性能要比组合成它们的单个分类器的结果好的多。如**图3.2** 集成学习所示，集成学习将M个学习得到的分类器，，……组合起来以创建一个改进的分类器。集成学习可以分为基于数据划分的方法（Data Partitioning Method）[29]和基于特征划分的方法（Attribute Partitioning Method）[30]，其中基于数据划分的方法有Bagging[31]、Boosting[32]，基于特征划分的方法有RS（Random Subspace） [11]。



图3.2 集成学习分类过程

（1）Bagging

Bagging（Boots trap Aggregation）是由Breiman在1996年提出，它的基本思想是可以多次从原始样本集中选取训练子集，每个训练样本对应一个分类器，使用投票的方式输出训练器的结果。训练中的样本是随机且有放回的选取，因此选取的样本子集允许多次被选中，通常情况下，每个基分类器的训练集中有63.2%的样本是重复的[20]。

（2）Boosting

Boosting算法思想最早来源于1984年的Valiant的PAC-Learning Model[33]。通常根据分类器的准确度将分类器分为弱分类器和强分类器，强分类器的准确率比较高，弱分类器的准确率仅比随机的猜测要好些。Boosting算法的核心思想是每个样本对应一个分类器，每个分类器的权重相等，在每次迭代中，按照一定的规则增加分类错误的样本的权重，减小分类正确样本的权重，使得下一次循环弱分类器能够集中精力去对错误的样本进行判断。Boosting具体的算法流程如图3.3。

（3）Random Subspace

Random Subspace是典型的基于特征划分的算法，RS方法主要是针对文本特征进行随机的处理，因此比较适用于文本分类等高维的分类问题。它的基本思想是随机的从特征样本集合中选取若干特征集，然后再这些特征集中训练得到分类器；最后使用组合得票的方式得到分类结果。

|  |
| --- |
| 1. 原始训练集输入，给每个样本赋予初始权重； 2. 将训练集输入已知的弱分类器，弱分类器对每个样本给出假设； 3. 更新训练集中各样本的权重； 4. 对此次的弱分类器给出权重； 5. 转到步骤2）直到循环达到一定次数或某度量标准符合要求； 6. 将弱分类器按其相应的权重加权组合形成强分类器。 |

图3.3 Boosting算法的基本过程

### 3.5.3 半监督学习与非均衡分类结合

针对电子商务中在线评论数据呈现非均衡的特点，本文提出了使用半监督学习与非均衡数据分类方法相结合的方式对在线评论进行有用性分析。在本文中所采用的半监督学习方法中的协同训练和自我训练算法，目前常用的针对非均衡数据分类方法可以分为基于抽样的方法和基于集成的学习方法。

|  |
| --- |
| **输入：**有标记样本集 M  无标记样本集 N  基础分类器 c  非均衡分类方法 CY  **过程：**在M中使用CY生成数据集d；  Loop for k 次循环  使用d训练一个分类器，只把视作的一部份；  使用d训练一个分类器，只把视作的一部份；  允许从N中标记p个有用的样本和n个没用的样本；  允许从N中标记p个有用的样本和n个没用的样本；  把这些标记的样本添加到d中；  **输出：** |

图3.4 基于非均衡分类方法与协同训练结合算法的过程

本文中以非均衡分类方法与协同训练相结合的方式生成CT-OS、CT-US、CT-SMOTE、CT-Bagging、CT-Boosting、CT-RS，具体的算法过程如图3.4；以非均衡分类方法与自我训练相结合的方式生成ST-OS、ST-US、ST-SMOTE、ST-Bagging、ST-Boosting、ST-RS，具体的算法过程如图3.5；实验中以SVM[34]作为实验的基本分类器，同时采用两倍实验交叉的方法对实验数据进行分析。使用半监督的学习方法与非均衡分类方法对实验数据进行有用性分析，首先，因为实验数据呈现出不均衡的特点，其次，基于半监督的学习方法可以少量标记样本数据，降低实验过程中的标记样本的人力、时间成本。

|  |
| --- |
| **输入：**已标记训练样本集L  未标记训练样本集U  利用L训练一个分类器h  非均衡分类方法 CY  **过程：** 在U中使用CY生成一个均衡数据集d；  从d中随机抽取u个样本组成训练池；  **迭代：** 使用h标记出中p个正样本和q个负样本；  将这些样本加入L，更新L；  从U中随机抽取(p+q)个样本来更新；  **停止条件：**训练池中已经没有未标记样本存在； |

图 3.5 基于非均衡分类方法与自我训练结合算法的过程

## 3.6 结果评价指标

在本论文实验的评价指标采用目前文本分类领域常用的评价指标：平均分类精度AA(Average Accuracy)指标和AUC(Area Under the Curve)指标**[25]**，其平均分类精度指标定义如公式(3-1)。

 (3-1)

其中，TP(True Positive)表示分类模型能够正确预测的正样本数，TN(True Negative)表示分类模型能够正确预测的负样本数，FP(False Positive)表示分类模型能够错误预测的正样本数，FN(False Negative)表示分类模型能够错误预测的负样本数。根据ROC(Receiver Operating Characteristic)曲线下的面积计算可以计算出AUC，本文将此方法作为结果分析的指标之一。

# 4 实验设计与结果分析

本段主要介绍了实验环境，实验数据获取的过程以及实验的具体流程，并对实验结果进行了整体分析、对实验中非均衡数据分类与半监督学习相结合方法做了对比分析，下面将详细介绍。

## 实验设计

### 4.1.1 实验环境

本论文中实验环境所采用的编程语言Java1.7，Java IDE使用的是MyEclipse 10.0，数据库使用MySql 5.6.19，使用Weka3.7作为文本分析的软件，以Visio2013作为流程图分析的工具，分词工具采用中国科学院计算机所编写的中文分词工具ICTCLAS。

### 4.1.2 实验数据

为了验证在电子商务基于半监督的学习对在线评论进行有用性分析的准确性和有效性。本文选取观众对电影的长评论作为在线评论的数据来源，本文之所以选用电影评论作为实验数据源是因为电影是人们公认的受口碑影响较大的商品[26]。电影短暂的上映周期，票房的收入随着季节、是否是周末波动，电影的更换速度也比较快，往往电影上映都根据其他电影的上映时间安排。由电影产生的在线评论也呈现出显著的特点，在电影上映的第一个星期六星期天、以及在贺岁档、暑期档、五一黄金周、十一黄金周这些黄金档期的期间，在线评论的数量快速增加，然后随着观看电影的人数近似以指数衰减，在线评论的数量也呈现出相似的规律。

本文中实验所使用的在线评论数据选自豆瓣电影的长评论，豆瓣网站提供书籍、电影、音乐等信息，用户既可以浏览其他用户发表的评论，还可以对其中好的评论点赞，也对这些作品发表长评论、短评论信息[36]。豆瓣网站还提供小组讨论、线下活动等多种服务功能。豆瓣致力于建立网络社区，帮助人们发现生活中有用、有趣的事物。数据库采用的是Mysql数据库。具体数据获取过程如图4.1。

图4.1 获取数据源

（1）使用Java语言编写程序从豆瓣电影网站抓取用户对发表的长评论。利用getHtmlContent()方法获取网页的内容，再使用正则表达式去匹配相关的内容，然后存入到数据库中。首先，豆瓣电影中每部电影的影评URL呈现一定的规律，本文采用顺序遍历的方式获取电影的URL；然后，根据某部电影的影评URL获取该部电影的长评论URL，好评数量，差评数量，将该部电影影评主页URL和该电影的长评论URL，好评数量，差评数量存入MySQL数据库；最后，由于在该影评URL主页，长评论部分被隐藏起来，所以本文需要访问该具体评论的URL，在该页面获取所有展开的评论内容，将该评论的URL和影评URL存入MySQL数据库中。

（2）为实验的在线评论分为有用和没用评论制定分类标准。根据好评数量在总的评论数量中占的比例是否大于等于0.6并且小于等于1.0，同时好评数量与差评数量之和大于0，将评论内容划分为有用和没用。最终结果共得到分类结果中包含730条有用评论和7496条没用评论。

（3）通过Java编程实现将MySQL数据库内的评论内容如果其status=true则逐条写入P文件夹下的每个txt中，如果status=false则逐条写入N文件夹下的每个txt中，文件名按数值逐渐递增。

（4）分词：编程实现调用中国科学院计算机所编写的中文分词工具ICTCLAS对评论本文语料进行分词，对原始文本的分词结果如表4.1。

（5）去停用词：本文主要利用哈工大信息检索中心提供的中文停用词表去除在线评论数据集中如副词、介词、感叹词、助词、拟声词等对没有提升分类性能的词语。

（6）特征选择，由于在线评论文本的特征数量庞大，本文选取了针对中文文本的5类100个特征，使用Java编程实现并对每个特征方法进行了单元测试。

（7）通过Java编程实现调用Weka的接口，将txt文档转变为分类工具Weka能够识别的ARFF格式，得到最终的分类数据。ARFF文件第一部分即第一行,以@relation开头，空格后面的豆瓣电影在线长评论表示数据的名称；第二部分以@atrribute标识，列举豆瓣电影在线长评论数据集中所有的属性，@atrribute后跟属性名，属性取值或属性类型；第三部分以@data开始，表示其后的数据为数据实例，每一行表示一个实例(共有100个实例)。

（8）编程调用Weka中的非均衡数据分类器与半监督学习相结合的方式对数据集进行分析，最终得到分类结果数据。

### 4.1.3 实验流程

本文使用Weka工具对语料库进行分类，其中基分类器采用Support Vector Machine(SVM)，分别选取Weka下SMO模块实现，选取Bagging模块具体实现Bagging，选取ADBoostMl模块模块具体实现Boosting和选取Random-subspace模块模块具体实现Random Subspace算法，对于取样方法OS、US、Smote则自行编程实现。为了提高实验结果的可信性和有效性，本实验采用了5次2倍交叉验证法，实验过程中分别以10%、20%、40%、60%、80%的比例分别标记样本，初始的样本池中的样本以1:10的比例从从标记样本和未标记样本的选取，然后将初始的数据集划分为2个大致相同的数据集，分别作为训练集和测试集，然后进行实验，在此步骤结束后将本次实验的测试集作为下一次实验的训练集，将本次实验的训练集作为下一次实验的测试集。本次实验结束后再以1:10的比例从各分类样本中选取数据加入到样本池中标记数据，然后进行下一次实验，具体的实验流程如图4.2。



图4.2 实验流程

## 4.2 实验结果分析

为了验证本实验所提出的基于半监督学习与非均衡数据分类方法相结合的方式对在线评论进行有用性分析的有效性，本文将实验分为结构特征和混合特征两部分以提高分类器分类精度，验证实验的可行性和有效性。本段主要介绍了基于结构特征的实验结果统计分析和基于混合策略的实验结果统计分析，现详细的介绍。

### 4.2.1 实验结果整体分析

在本实验中本文选取SVM作为分类模型标准。ST表示自我训练算法(Self-Training)，CT表示协同训练算法(Co-Training)，ST-OS、ST-US、ST-Smote、ST-Bagging、ST-Boosting、ST-RS分别表示ST与OS、ST与US、ST与Smote、ST与Bagging、ST与Boosting、ST与RS相结合的方法，CT-OS、CT-US、CT-Smote、CT-Bagging、CT-Boosting、CT-RS分别表示CT与OS、CT与US、CT与Smote、CT与Bagging、CT与Boosting、CT与RS相结合的方法。

基于结构特征的实验结果统计，详见表4.2，在标记比率10%、20%、40%的标记比率，CT-RS在AA、AUC指标下都取得了最好的结果，在60%标记比率时，AA指标下也是CT-RS取得了最好的结果，AUC指标下CT-OS取得了最好的结果，在80%标记比率时，CT-Boosting在AA指标下取得了最好的结果，CT-OS在AUC指标下取得了最好的结果。

在基于混合策略结构特征的实验结果统计中，如表4.3，在10%标记比率时，在AA指标下CT-RS取得了最好的结果，在AUC指标时CT-OS取得了最好的结果；在20%标记比率时，在AA指标下CT-RS取得了最好的结果，在AUC指标时ST-RS取得了最好的结果；在40%标记比率时，在AA指标下ST-RS取得了最好的结果，在AUC指标时CT-Smote取得了最好的结果；在60%标记比率时，在AA指标下ST-RS取得了最好的，在AUC指标时CT-RS取得了最好的结果；在80%标记比率时，在AA指标下ST-RS取得了最好的结果，在AUC指标时CT-Bagging取得了最好的结果。

如表4.1、表4.2所示，在同时考虑仅基于结构特征和混合策略时，在使用半监督学习方法与分均衡数据分类方法相结合的方式中，CT-Bagging、CT-Boosting、CT-RS、ST-RS、CT、ST在10%、20%、40%、60%、80%标记比例下，AA和AUC指标均取得了比SVM更好的结果。

表4.1 基于结构特征的实验结果

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **标记比例**  **方法** | **10%** | | **20%** | | **40%** | | **60%** | | **80%** | |
| **AA(%)** | **AUC** | **AA(%)** | **AUC** | **AA(%)** | **AUC** | **AA(%)** | **AUC** | **AA(%)** | **AUC** |
| **SVM** | 87.44 | 0.6611 | 90.08 | 0.6831 | 90.92 | 0.6760 | 91.09 | 0.6753 | 91.07 | 0.6547 |
| **OS** | 75.87 | 0.6919 | 73.32 | 0.7135 | 73.09 | 0.7419 | 73.68 | 0.7488 | 73.67 | 0.7567 |
| **US** | 56.67 | 0.6375 | 59.96 | 0.6629 | 64.94 | 0.7166 | 66.44 | 0.7324 | 66.46 | 0.7347 |
| **Smote** | 86.13 | 0.6837 | 89.01 | 0.7113 | 89.82 | 0.7238 | 90.35 | 0.7408 | 90.65 | 0.7366 |
| **Bagging** | 87.84 | 0.6661 | 89.91 | 0.6967 | 90.88 | 0.7115 | 91.06 | 0.7154 | 91.07 | 0.7150 |
| **Boosting** | 85.12 | 0.6351 | 87.47 | 0.6780 | 89.78 | 0.7041 | 90.62 | 0.7195 | 90.90 | 0.7296 |
| **RS** | 90.36 | 0.6870 | 90.93 | 0.6912 | 91.06 | 0.7115 | 91.11 | 0.6973 | 91.10 | 0.7181 |
| **ST** | 88.76 | 0.7088 | 90.33 | 0.7339 | 90.95 | 0.7355 | 91.11 | 0.7277 | 91.11 | 0.7238 |
| **ST-OS** | 75.99 | 0.7166 | 73.71 | 0.7289 | 73.80 | 0.7477 | 74.10 | 0.7521 | 73.76 | 0.7584 |
| **ST-US** | 58.14 | 0.6656 | 60.45 | 0.6908 | 65.50 | 0.7276 | 67.30 | 0.7401 | 67.07 | 0.7438 |
| **ST-Smote** | 86.92 | 0.7194 | 89.27 | 0.7456 | 89.98 | 0.7512 | 90.39 | 0.7533 | 90.66 | 0.7516 |
| **ST-Bagging** | 88.67 | 0.7183 | 90.12 | 0.7363 | 90.90 | 0.7392 | 91.08 | 0.7400 | 91.13 | 0.7407 |
| **ST-Boosting** | 85.24 | 0.6562 | 87.84 | 0.6860 | 90.13 | 0.7090 | 90.71 | 0.7229 | 90.99 | 0.7342 |
| **ST-RS** | 90.59 | 0.7448 | 91.02 | 0.7571 | 91.10 | 0.7592 | 91.16 | 0.7544 | 91.15 | 0.7596 |
| **CT** | 90.27 | 0.7356 | 90.82 | 0.7487 | 91.07 | 0.7539 | 91.13 | 0.7328 | 91.13 | 0.7326 |
| **CT-OS** | 76.01 | 0.7308 | 74.66 | 0.7481 | 75.26 | 0.7623 | 75.66 | **0.7639** | 75.45 | **0.7682** |
| **CT-US** | 57.99 | 0.6692 | 62.32 | 0.6923 | 66.87 | 0.7347 | 69.22 | 0.7505 | 69.38 | 0.7526 |
| **CT-Smote** | 89.20 | 0.7383 | 90.30 | 0.7571 | 90.71 | 0.7615 | 90.87 | 0.7639 | 90.99 | 0.7594 |
| **CT-Bagging** | 90.55 | 0.7391 | 91.04 | 0.7502 | 91.11 | 0.7559 | 91.15 | 0.7460 | 91.14 | 0.7500 |
| **CT-Boosting** | 89.75 | 0.6990 | 90.71 | 0.7269 | 91.03 | 0.7410 | 91.14 | 0.7509 | **91.17** | 0.7544 |
| **CT-RS** | **91.13** | **0.7450** | **91.14** | **0.7605** | **91.16** | **0.7662** | **91.16** | 0.7624 | 91.16 | 0.7663 |

表4.2 基于混合策略的实验结果

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **标记比例**  **方法** | **0.1** | | **0.2** | | **0.4** | | **0.6** | | **0.8** | |
| **AA(%)** | **AUC** | **AA(%)** | **AUC** | **AA(%)** | **AUC** | **AA(%)** | **AUC** | **AA(%)** | **AUC** |
| **SVM** | 86.51 | 0.8100 | 87.82 | 0.8332 | 90.03 | 0.8502 | 90.53 | 0.8512 | 91.15 | 0.8624 |
| **OS** | 89.21 | 0.6979 | 88.59 | 0.8289 | 86.89 | 0.8355 | 85.92 | 0.8450 | 85.77 | 0.8603 |
| **US** | 72.63 | 0.8220 | 75.60 | 0.8408 | 79.11 | 0.8589 | 79.66 | 0.8738 | 80.48 | 0.8831 |
| **Smote** | 87.09 | 0.8152 | 88.48 | 0.8373 | 89.59 | 0.8576 | 89.89 | 0.8550 | 90.59 | 0.8742 |
| **Bagging** | 86.56 | 0.8386 | 89.12 | 0.8606 | 90.52 | 0.8775 | 90.80 | 0.8836 | 91.30 | 0.8904 |
| **Boosting** | 87.52 | 0.7344 | 88.08 | 0.7922 | 88.52 | 0.8220 | 89.18 | 0.8442 | 89.90 | 0.8584 |
| **RS** | 90.15 | 0.8558 | 90.87 | 0.8792 | 91.55 | 0.8945 | 91.46 | 0.9012 | 91.67 | 0.9064 |
| **ST** | 86.88 | 0.8207 | 88.15 | 0.8447 | 90.08 | 0.8565 | 90.59 | 0.8554 | 91.18 | 0.8647 |
| **ST-OS** | 89.11 | 0.7958 | 88.58 | 0.8389 | 87.63 | 0.8433 | 86.89 | 0.8500 | 86.88 | 0.8615 |
| **ST-US** | 77.62 | 0.8271 | 78.88 | 0.8489 | 80.09 | 0.8642 | 80.16 | 0.8769 | 80.64 | 0.8841 |
| **ST-Smote** | 85.42 | 0.8241 | 86.38 | 0.8504 | 87.73 | 0.8635 | 87.98 | 0.8622 | 88.63 | 0.8773 |
| **ST-Bagging** | 87.71 | 0.8458 | 89.34 | 0.8685 | 90.65 | 0.8826 | 90.91 | 0.8862 | 91.41 | 0.8950 |
| **ST-Boosting** | 87.60 | 0.7804 | 88.46 | 0.8131 | 89.18 | 0.8405 | 89.76 | 0.8559 | 90.34 | 0.8718 |
| **ST-RS** | 90.65 | 0.8947 | 91.23 | **0.9057** | **91.77** | 0.9066 | **91.77** | 0.9092 | **91.96** | 0.9130 |
| **CT** | 90.40 | 0.8790 | 91.15 | 0.8932 | 91.66 | 0.9008 | 91.63 | 0.9015 | 91.84 | 0.9036 |
| **CT-OS** | 88.75 | **0.8957** | 87.85 | 0.9030 | 86.16 | 0.9005 | 85.63 | 0.9035 | 85.58 | 0.9053 |
| **CT-US** | 74.91 | 0.8552 | 79.99 | 0.8892 | 81.85 | 0.8981 | 82.67 | 0.9063 | 83.20 | 0.9099 |
| **CT-Smote** | 90.05 | 0.8828 | 90.75 | 0.8970 | 91.34 | **0.9072** | 91.45 | 0.9076 | 91.64 | 0.9096 |
| **CT-Bagging** | 90.38 | 0.8839 | 91.09 | 0.8979 | 91.63 | 0.9071 | 91.65 | **0.9098** | 91.86 | **0.9135** |
| **CT-Boosting** | 89.53 | 0.8609 | 90.42 | 0.8819 | 91.07 | 0.8943 | 91.41 | 0.8970 | 91.74 | 0.9003 |
| **CT-RS** | **91.04** | 0.8901 | **91.31** | 0.9024 | 91.50 | 0.9054 | 91.62 | 0.9088 | 91.71 | 0.9119 |

### 4.2.2 基于结构特征的实验结果统计分析

本段展示和分析了基于结构特征在标记比例为10%、20%、40%、60%、80%时，AA和AUC指标下，基本分类器、自我训练、协同训练、基于自我训练与非均衡数据分类方法相结合、基于协同训练与非均衡数据分类方法相结合等实验结果，以及这些相结合方式相对于分类器SVM的精度提高百分比。

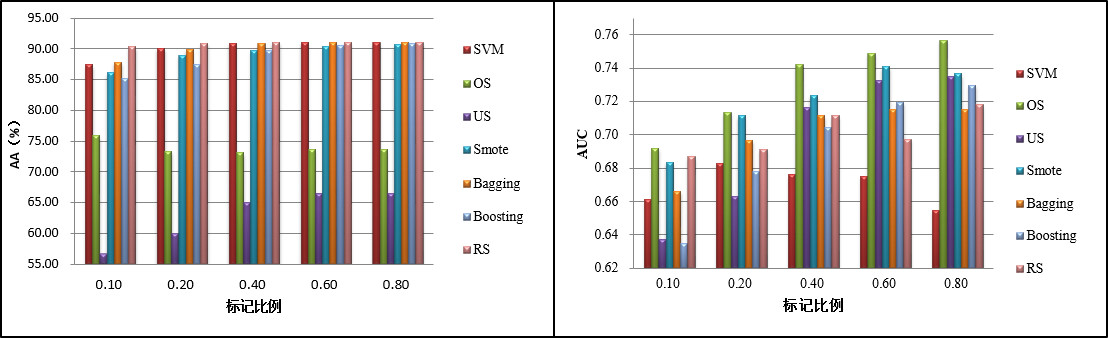


图4.3 基于结构特征，非均衡分类结果分析

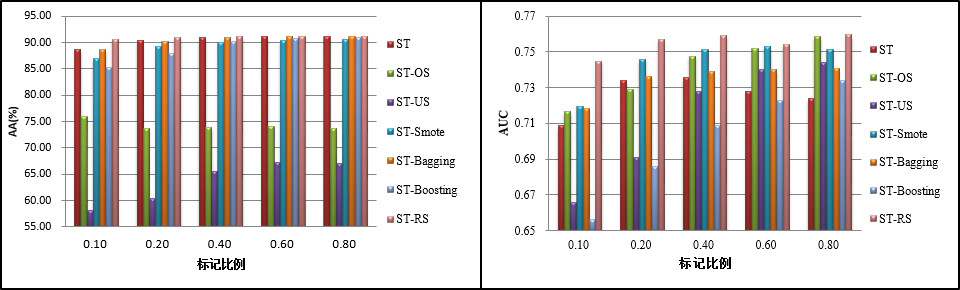


图4.4 使用结构特征，基于自我训练的非均衡分类方法的结果对比分析

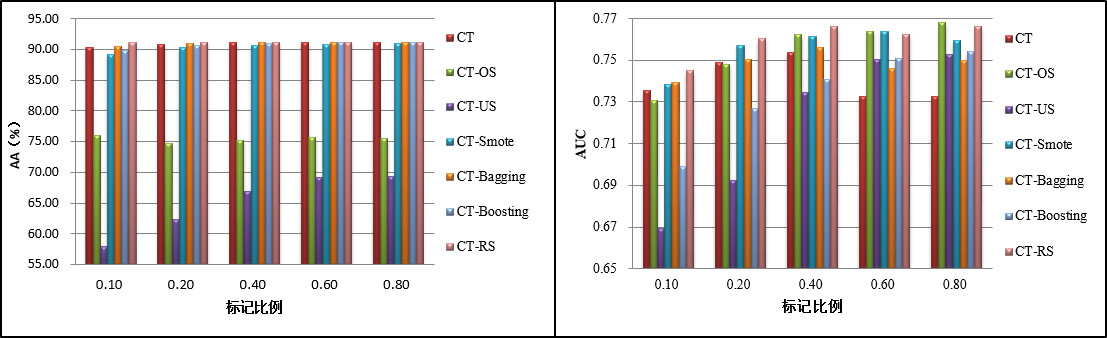


图4.5 使用结构特征，基于协同训练的非均衡分类方法的结果对比分析

如图4.3所示，在仅使用结构特征的实验中AA 指标下，不同标记比率对于基分类器的影响中，在不同的标记比率中SVM、Smote、Bagging、Boosting、RS分类方法结果大致相等，且明显优于US、OS，而OS又优于US；在AUC 指标下，不同标记比率对于基分类器的影响中，随着标记比率的增加SVM、Smote、Bagging、Boosting、RS、US、OS的分类结果总体上呈现上升的趋势。在不同标记比例下，Smote和集成学习的方法取得了和SVM大致相等的AA值，OS、Smote、Bagging、RS取得了比SVM更好的AUC值，并且随着标记比例的增加抽样学习的方法和集成学习的方法取得的AUC值不断增加，标记比例越高取得了比SVM更好的值。

如图4.4所示，在本实验中使用结构特征，基于自我训练方法与非均衡相结合的方式中，在AA 指标下，不同标记比率对于分类器的影响中，在不同的标记比率中ST、ST-Smote、ST-Bagging、ST-Boosting、ST-RS分类方法明显优于ST-OS、ST-US，而ST-OS又优于ST-US；在AUC 指标下，不同标记比率对于ST、ST-OS、ST-US、ST-Smote、ST-Bagging、ST-Boosting、ST-RS分类方法的影响相对较小。

如图4.5所示，在本实验中使用结构特征，基于协同训练方法与非均衡相结合的方式中，在AA 指标下，不同标记比率对于分类器的影响中，在不同的标记比率中CT、CT-Smote、CT-Bagging、CT-Boosting、CT-RS分类方法明显优于CT-OS、CT-US，而CT-OS又优于CT-US，但是他们之间的差距随着标记比率的增加呈现逐渐缩小的趋势；在AUC 指标下，不同标记比率对于CT、CT-Smote、CT-Bagging、CT-Boosting、CT-RS、CT-OS、CT-US分类方法的影响相对较小。

不同的半监督学习方法相对于基础分类器SVM的分类精度和AUC提高的百分比的计算方法，如公式(4-1)、(4-2)。

 (4-1)

 (4-2)

如图4.6、图4.7可以得出由以上实验结果可以看出，在本实验结果中，在自我训练与基于非均衡数据分类相结合的方式中，ST-RS在10 %、20%、40%、60%、80%的标记比率，AA和AUC指标下都取得了最好的结果。在协同训练与基于非均衡数据分类相结合的方式中，CT-RS在10 %、20%、20%、40%、60%、80%的标记比率，AA和AUC指标下都取得了最好的20%、40%的标记比率，AA和AUC指标下都取得了最好的结果；在60%的标记比率下，在AA指标下，CT-RS也取得了最好的结果；在AUC指标下，CT-OS和CT-SMOTE同时取得了最好的效果；在80%的标记比率下，在AA指标下，CT-Boosting取得了最好的结果；在AUC指标下，CT-OS取得了最好的效果。

如图4.8所示，在仅使用结构特征的实验中AA 指标下，不同标记比率对于协同训练的影响中，随着标记比率的变化CT、CT-OS、CT-Smote、CT-Bagging、CT-Boosting、CT-RS呈现较小的变化；CT-US随着标记比率增加逐渐增加，然后慢慢变得平稳；在AUC 指标下，不同标记比率对于协同训练的影响中，随着标记比率的变化CT-OS、CT-Smote、CT-Bagging、CT-RS呈现先上升后稳定的变化趋势；CT-Bagging 、CT随着标记比率增加逐渐增加，然后下降再平稳的趋势；CT-Boosting、CT-US随着标记比率增加逐渐上升的趋势。

图4.6 使用结构特征，不同分类方法的分类精度提高百分比(AA)

图4.7 使用结构特征，不同分类方法的分类精度提高百分比(AUC)

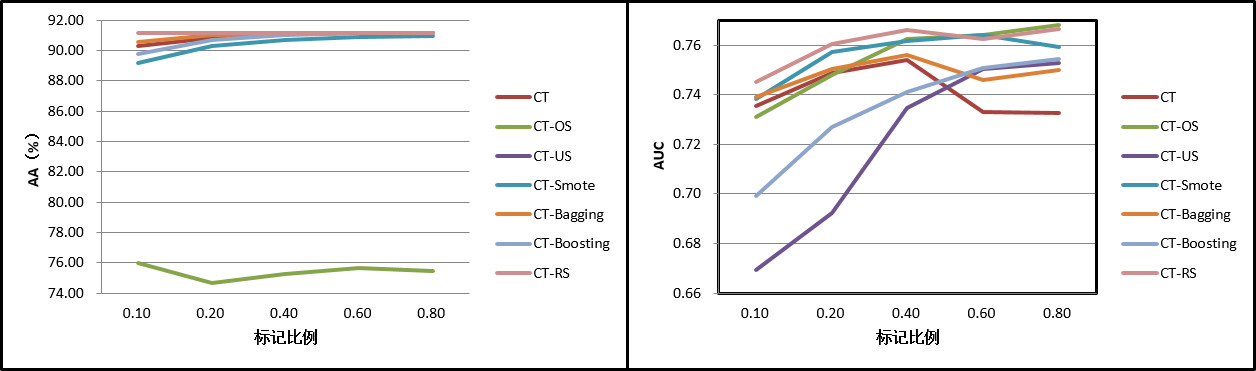


图4.8 使用结构特征，不同标记比例对于协同训练的影响

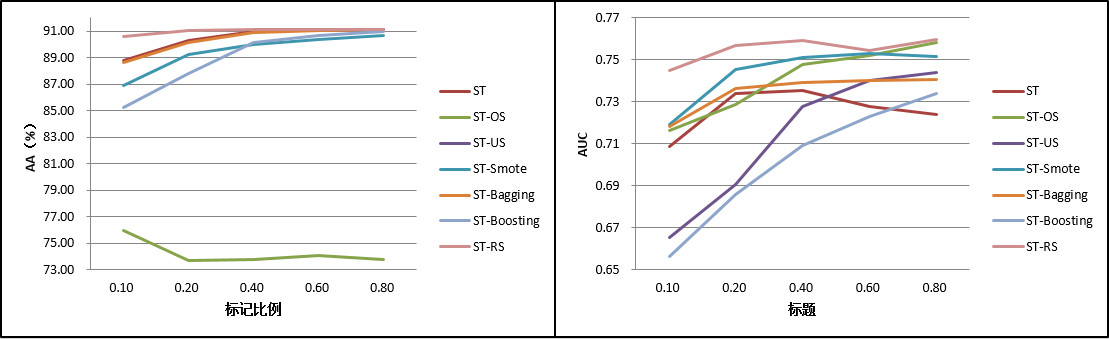


图4.9 使用结构特征，不同标记比例对于自我训练的影响

如图4.9所示，在使用混合策略的实验中AA 指标下，不同标记比率对于自我训练的影响中，随着标记比率的增加ST-RS没有表现出明显的变化；ST、ST-Smote、ST-Boosting、ST-Bagging随着标记比率的增加先上升，然后变得比较稳定；ST-OS随着标记比率的增加，先小幅下降，然后呈现出比较稳定；ST-US随着标记比率的增加，先增加，然后呈稳定的趋势；在AUC 指标下，不同标记比率对于自我训练的影响中，随着标记比率的增加ST呈现出先增加然后缓慢降低的趋势；ST-OS、ST-Smote、ST-Bagging、ST-RS随着标记比率的增加先上升，然后变得比较稳定；ST-OS随着标记比率的增加，先小幅下降，然后呈现出比较稳定；ST-US、ST-Boosting随着标记比率的增加逐步增加的趋势。

### 4.2.3 基于混合策略的实验结果统计分析

本段展示和分析了基于混合策略在标记比例为10%、20%、40%、60%、80%时，AA和AUC指标下，基本分类器、自我训练、协同训练、基于自我训练与非均衡数据分类方法相结合、基于协同训练与非均衡数据分类方法相结合等实验结果，以及这些相结合方式相对于分类器SVM的精度提高百分比。

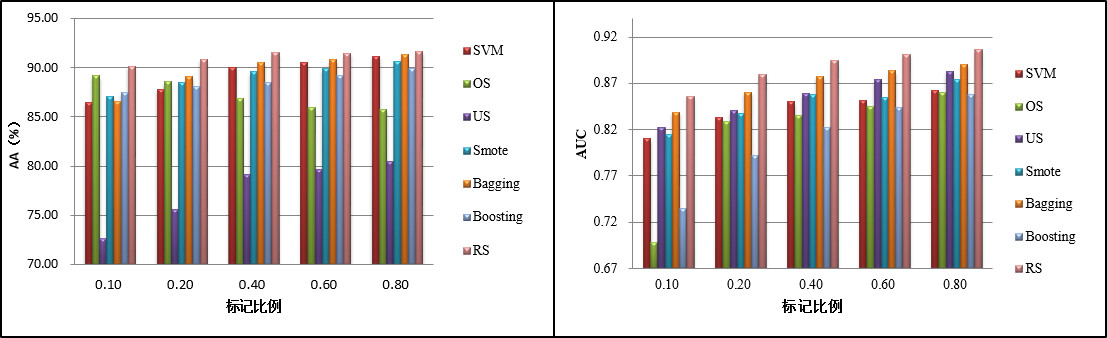


图4.10 基于混合策略下，非均衡分类方法结果分析

如图4.10所示，在使用混合策略的实验中AA 指标下，不同标记比率对于基分类器的影响中，在不同的标记比率中SVM、Smote、Bagging、Boosting、RS分类方法明显优于US、OS，而OS又优于US，但随着标记比率的增加他们之间的差距逐渐降低；在AUC 指标下，不同标记比率对于基分类器的影响中，随着标记比率的增加SVM、Smote、Bagging、Boosting、RS分类方法分类效果整体上处于上升趋势。

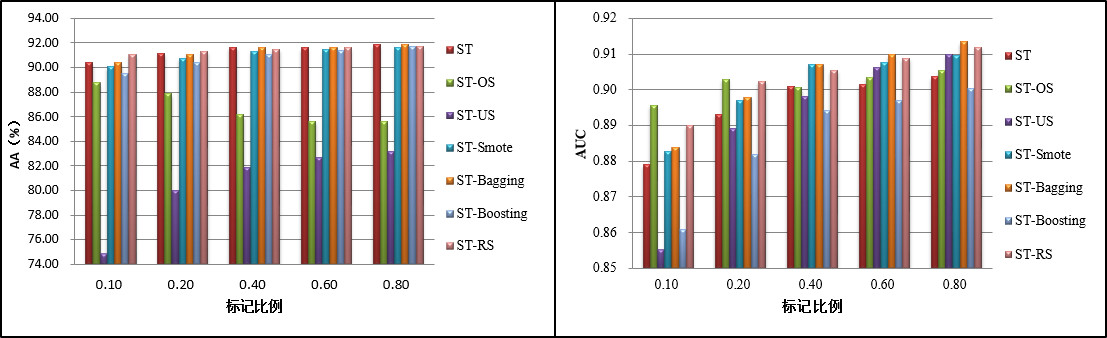


图4. 11 基于混合策略下，不同标记比例下基于自我训练的分类方法的影响

如图4.11所示，在使用混合策略的实验中，基于自我训练方法与非均衡相结合的方式在AA 指标下，不同标记比率对于分类器的影响中，在不同的标记比率中CT、CT-Smote、CT-Bagging、CT-Boosting、CT-RS分类方法明显优于CT-OS、CT-US，而CT-OS又优于CT-US，但是他们之间的差距随着标记比率的增加呈现逐渐缩小的趋势；在AUC 指标下，不同标记比率对于分类器的影响中，在不同的标记比率中CT、CT-Smote、CT-Bagging、CT-Boosting、CT-RS、CT-OS、CT-US之间的差距并不是特别大，而且随着标记比率的增加他们呈现缓慢上升的趋势。

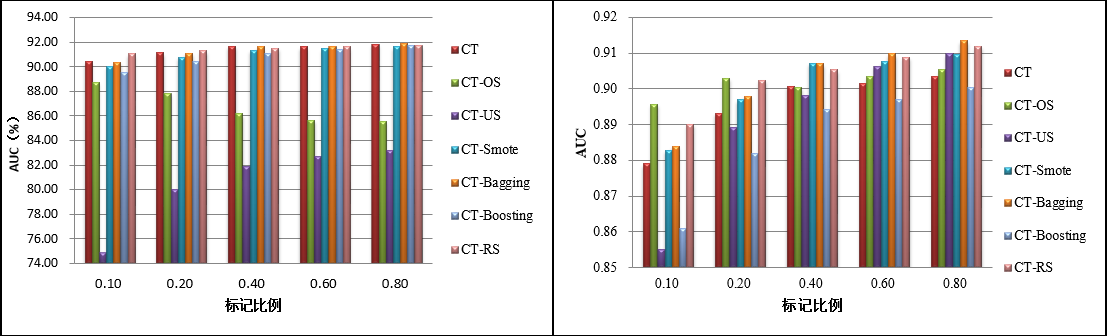


图4.12 基于混合策略下，不同标记比例下基于协同训练的分类方法的影响

如图4.12所示，在使用混合策略的实验中，基于协同训练方法与非均衡相结合的方式在AA 指标下，不同标记比率对于分类器的影响中，在不同的标记比率中CT、CT-Smote、CT-Bagging、CT-Boosting、CT-RS分类方法明显优于CT-OS、CT-US，而CT-OS又优于CT-US，但是他们之间的差距随着标记比率的增加呈现明显的缩小的趋势；在AUC 指标下，随着标记比率的增加不同的分类方法的分类结果总体上呈现缓慢增加的趋势。实验指标AA和AUC与基于混合特征中的实验评价指标的计算方式如公式(4-1)、(4-2)相同。

图4.13 基于混合策略下，不同分类方法的平均分类精度提高百分比

图4.14基于混合策略下，不同分类方法AUC提高结果百分比

如图4.13、图4.14所示，在自我训练与基于非均衡数据分类相结合的方式中，ST-RS在10 %、20%、40%、60%、80%的标记比率，AA和AUC指标下都取得了最好的结果。在协同训练与基于非均衡数据分类相结合的方式中，在10%的标记比率时，在AA指标下，CT-RS也取得了最好的结果；在AUC指标下，CT-OS取得了最好的效果；在20%的标记比率时，在AA指标下，CT-RS取得了最好的结果；在AUC指标下，CT-OS取得了最好的效果；在40%的标记比率时，在AA指标下，CT取得了最好的结果；在AUC指标下，CT-Smote取得了最好的效果；在60%和80%的标记比率时，在AA和AUC指标时，CT-Bagging均取得了最好的结果。

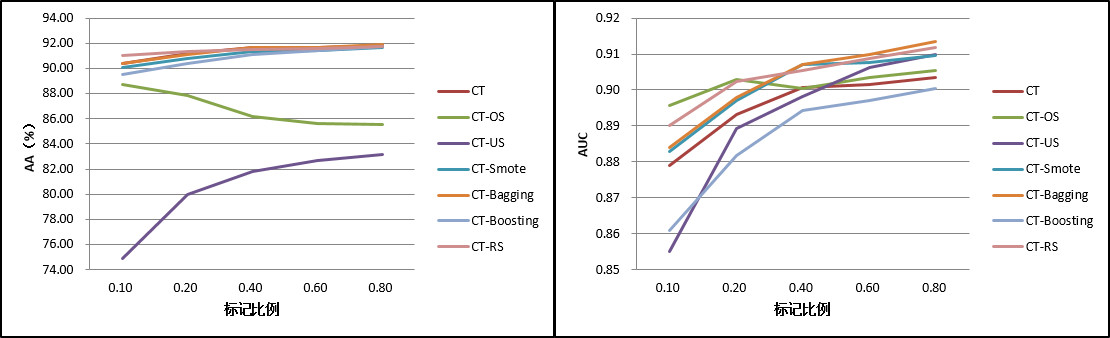


图4.15基于混合策略下，在不同的比例下基于协同训练的分类方法的影响

如图4.15所示，在基于结构特征的实验中在AA 指标下，不同标记比率对于协同训练的影响中，随着标记比率的变化CT、CT-Smote、CT-Bagging 、CT-Boosting、CT-RS呈现先缓慢上升后稳定的变化趋势；CT-US随着标记比率增加逐渐增加，然后逐渐上升的趋势；CT-OS随着标记比率增加缓慢下降的趋势；在AUC 指标下，不同标记比率对于协同训练的影响中，随着标记比率的变化CT、CT-OS、CT-US、CT-Smote、CT-Bagging 、CT-Boosting、CT-RS呈现先缓慢上升后稳定的变化趋势。

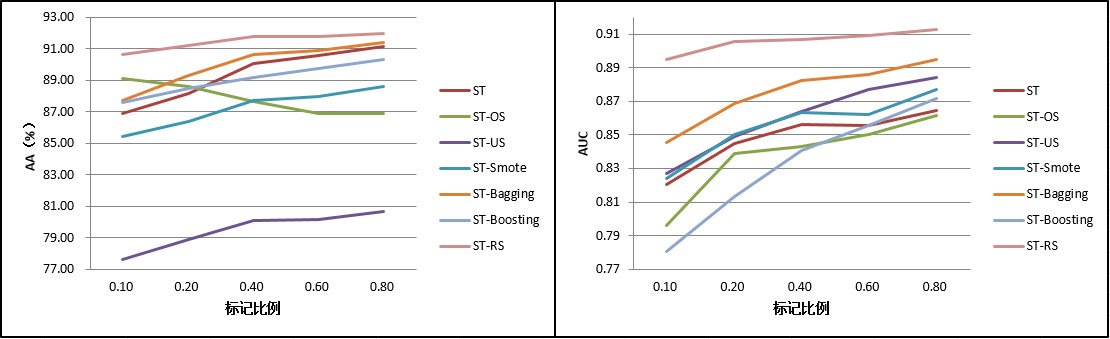


图4.16基于混合策略下，在不同的比例下基于自我训练的分类方法的影响

如图4.16所示，在基于结构特征的实验中在AA 指标下，不同标记比率对于自我训练的影响中，随着标记比率的变化ST、ST-OS、ST-US、ST-Bagging、 ST-Boosting、ST-RS呈现先缓慢上升后稳定的变化趋势；随着标记比率的变化ST-Smote呈现出先缓慢下降再稳定的趋势；在AUC 指标下，不同标记比率对于自我训练的影响中，随着标记比率的变化ST、ST-OS、ST-US、ST-Smote、ST-Bagging、ST-Boosting、ST-RS呈现先缓慢上升的变化趋势。

### 4.2.4 结构特征下不同添加样本对实验结果统计分析

本论文研究在10%、20%、40%、60%、80%标记比例下，以1:10比例从有用和没用的数据中选取样本添加进样本池，具体添加数量分别选取（p=1,n=10）、（p=2,n=20）、（p=3,n=30）、（p=4,n=40）、（p=5,n=50）。本段在仅使用结构特征进行实验，最终生成结果，下面对统计结果进行分析。

|  |
| --- |
|  |
|  |

图4.17 基于结构特征，不同添加样本对于相结合方法的影响（10%标记比率

如图4.17所示，在10%标记比例下基于半监督学习与非均衡数据分类的结合的方法中，在AA指标下，随着样本添加的改变各个方法表现的比较稳定，在AUC结果下，随着样本的增加，ST-US、CT-Smote、CT-Bagging出现结果稍微下降趋势；ST-Bagging表现略微的上升，其它结果在总体上保持稳定。

|  |
| --- |
|  |
|  |

图4.18 基于结构特征，不同添加样本对于相结合方法的影响（20%标记比率）

如图4.18所示，在20%标记比例下基于半监督学习与非均衡数据分类的结合的方法在AA、AUC指标下结果在总体上保持稳定。

|  |
| --- |
|  |
|  |

图4.19 基于结构特征，不同添加样本对于相结合方法的影响（40%标记比率）

如图4.19所示在40%标记比例下基于半监督学习与非均衡数据分类的结合的方法中，在AA指标下，随着样本添加的改变各个方法表现的比较稳定，在AUC结果下，随着样本的增加，ST-Boosting、ST-US、ST-Bagging、CT-US、CT-Bagging出现结果稍微下降趋势。

|  |
| --- |
|  |
|  |

图4.20 基于结构特征，不同添加样本对于相结合方法的影响（60%标记比率）

如图4.20所示，在60%标记比例下基于半监督学习与非均衡数据分类的结合的方法中，在AA指标下，随着样本添加的改变各个方法表现的比较稳定，在AUC结果下，随着样本的增加，ST-US、CT-US、CT-Bagging出现结果稍微下降趋势。

|  |
| --- |
|  |
|  |

图4.21 基于结构特征，不同添加样本对于相结合方法的影响（80%标记比率）

如图4.21所示，在80%标记比例下基于半监督学习与非均衡数据分类的结合的方法在AA、AUC指标下结果在总体上保持稳定。

### 4.2.5 混合策略下不同添加样本对实验结果统计分析

如图4.22、图4.23、图4.24、图4.25、图4.26所示，在10%、20%、40%、60%、80%标记比例下，基于半监督学习与非均衡数据分类的结合的方法在AA、AUC指标下结果在总体上保持稳定。表明混合策略比仅基于结构特征的策略更可靠，稳定；而且使用协同训练与非均衡数据的分类相结合的方法比使用自我训练和非均衡数据分类的方法取得了更好的结果，尤其在AUC指标下。

|  |
| --- |
|  |
|  |

图4.22 基于混合策略，不同添加样本对于相结合方法的影响（10%标记比率）

|  |
| --- |
|  |
|  |

图4.23基于混合策略，不同添加样本对于相结合方法的影响（20%标记比率）

|  |
| --- |
|  |
|  |

图4.24基于混合策略，不同添加样本对于相结合方法的影响（40%标记比率）

|  |
| --- |
|  |
|  |

图4.25基于混合策略，不同添加样本对于相结合方法的影响（60%标记比率）

|  |
| --- |
|  |
|  |

图4.26基于混合策略，不同添加样本对于相结合方法的影响（80%标记比率）

### 4.2.6 混合策略和结构特征的对比分析

由表4.4分析可得，在混合策略和仅结构特征的实验中AA指标下，混合策略中有13个方法比仅使用结构特征的方法取得了较好的结果；在AA指标下在所有的方法均取得了比基于结构特征的方法更好的实验结果；实验结果可以分析出使用混合策略的方式比仅使用结构特征的方法实验结果更精确。

表4.3 混合策略与结构特征实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **评价指标**  **实验方法** | **AA** | | **AUC** | |
| **混合策略** | **结构特征** | **混合策略** | **结构特征** |
| **SVM** | 89.21 | **90.12** | **0.8414** | 0.6700 |
| **OS** | **87.28** | 73.93 | **0.8135** | 0.7306 |
| **US** | **77.50** | 62.89 | **0.8557** | 0.6968 |
| **Smote** | 89.13 | **89.19** | **0.8479** | 0.7192 |
| **Bagging** | 89.66 | **90.15** | **0.8701** | 0.7009 |
| **Boosting** | 88.64 | **88.78** | **0.8102** | 0.6933 |
| **RS** | **91.14** | 90.91 | **0.8874** | 0.7010 |
| **ST** | 89.38 | **90.45** | **0.8484** | 0.7259 |
| **ST-OS** | **87.82** | 74.27 | **0.8379** | 0.7407 |
| **ST-US** | **79.48** | 63.69 | **0.8602** | 0.7136 |
| **ST-Smote** | 87.23 | **89.44** | **0.8555** | 0.7442 |
| **ST-Bagging** | 90.00 | **90.38** | **0.8756** | 0.7349 |
| **ST-Boosting** | **89.07** | 88.98 | **0.8323** | 0.7017 |
| **ST-RS** | **91.48** | 91.00 | **0.9058** | 0.7550 |
| **CT** | **91.34** | 90.88 | **0.8956** | 0.7407 |
| **CT-OS** | **86.79** | 75.41 | **0.9016** | 0.7547 |
| **CT-US** | **80.52** | 65.16 | **0.8917** | 0.7199 |
| **CT-Smote** | **91.05** | 90.41 | **0.9008** | 0.7560 |
| **CT-Bagging** | **91.32** | 91.00 | **0.9024** | 0.7482 |
| **CT-Boosting** | **90.83** | 90.76 | **0.8869** | 0.7344 |
| **CT-RS** | **91.44** | 91.15 | **0.9037** | 0.7601 |

# 结 论

电子商务的快速发展与普及，消费者的日常活动中越来越依靠互联网，消费者不仅将自己的使用体验发布在网络上，也将在线评论作为自己的参考信息，但是在线评论数量巨大、增长速度快并且质量良莠不齐，消费者需要识别出有用的在线评论。由于获取大量标记样本比较困难，却可以轻易获取大量无标记的样本，并且这些数据呈现非均衡的特点，本文中利用基于半监督学习方法和非均衡数据分类方法相结合的方式对在线评论做有用性分析。

本文首先使用Java编写在线评论数据抓取程序，其次对初始数据进行预处理，编程实现100中文文本特征，然后在实验中以半监督学习与非均衡分类方法相结合的方式与半监督学习方法、分均衡数据分类方法进行结果对比，最终得到的实验结果证明了使用本文提出基于半监督学习与在线评论相结合的方式对在线评论进行有用性分析是可行有效的。

当然本文也存在着不足，本文对在线评论的数据源只抓取了豆瓣电影的长评论，没有做到实验数据源的多样性；其次，本文抓取的评论数据是长评论，并没有抓取段评论，短评论数量比长评论更多，并且包含了丰富的隐藏信息；最后，在对数据源进行有用和没用的信息筛选的时候，将只有有用并且评论数足够的影评划分类别为没用，而实际上这部分信息可能是有用的。

# 参考文献

[1] 中国互联网络信息中心.第35次中国互联网络发展状况统计报告[R].北京, 2015.

[2] 杨娟.大学生写博动机调查分析:以中国大学生博客圈为例[J].赤峰学院学报: 哲学社会科学版, 2010, 16(11): 197—199.

[3] 梁吉业,高嘉伟,常瑜.半监督学习研究进展[J].山西大学学报(自然科学版),2009,04:528-534.

[4] 郭国庆,陈凯,何飞.消费者在线评论可信度的影响因素研究[J].当代经济管理,2010,10:17-23.

[5] 中国互联网络信息中心.2013年中国网络购物市场研究报告[R].北京, 2013.

[6] 张建欣.基于半监督学习的在线评论挖掘应用[J].计算机光盘软件与应用,2012,20:73-74.

[7] 薛为民,陆玉昌.文本挖掘技术研究[J].北京联合大学学报(自然科学版),2005,04:59-63.

[8] Jiawei Han,Micheline Kamber.数据挖掘技术和概念.北京: 机械工业出版社,2002.

[9] 诺志群,张国煊.文本挖掘研究进展[J],模式识别与人工智能,2005.05,第18卷 第1期: 3-4.

[10] 杨斌.中文文本数据挖掘研究[D].湘潭大学,2002.

[11] Ho, T.K.The random subspace method for constructing decision forests [J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 832-844.

[12] 王国才.朴素贝叶斯分类器的研究与应用[D].重庆交通大学,2010.

[13] 张庆国,张宏伟,张君玉.一种基于k最近邻的快速文本分类方法[J].中国科学院研究生院学报,2005,05:554-559.

[14] 郭茂.基于类中心向量的文本分类模型研究与实现[D].大连理工大学,2010.

[15] 刘华煜.基于支持向量机的机器学习研究[D].大庆石油学院,2005.

[16] 李荣陆,王建会,陈晓云,陶晓鹏,胡运发.使用最大熵模型进行中文文本分类[J].计算机研究与发展,2005,01:94-101.

[17] 李欣.基于神经网络的数据挖掘方法研究[D].大庆石油学院,2003.

[18] 冯少荣.决策树算法的研究与改进[J].厦门大学学报(自然科学版),2007,04:496-500.

[19] 陈莉平.基于关联规则的数据挖掘算法研究[D].江南大学,2008.

[20] 唐焕玲.基于半监督与集成学习的文本分类方法.电子工业出版社.2013.8:7-8.

[21] Blum A, Mitchell T.Combining Labeled and Unlabeled Data with Co-Training [C].In: Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT’98).New York, USA: ACM, 1998: 92-100.

[22]Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang.A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques, JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 57(3):378–393, 2006

[23] Fiona, J, Tweedie, R, Harald, Baayen.How Variable May a Constant be? Measures of Lexical Richness in Perspective[J].Computers and the Humanities, 1998, (32): 323-352

[24] 钱云.非均衡数据分类算法若干应用研究[D].吉林大学,2014.

[25] He H,Garcia E A.Learning from imbalanced data[J].IEEE Transactions on Knowledge and Data Engineering,2009,21(9): 1263-1284.

[26] Chawla N V, Japkowicz N, Kotcz A.Editorial: special issue on learning from imbalanced data sets[J].ACM SIGKDD Explorations Newsletter,2004, 6(1):1-6.

[27] Chawla N V, Bowyer K W,Hall L 0, et a1.SMOTE: synthetic minority over-sampling technique[J].Journal of Artificial Intelligence Research,2002,16: 321-357.

[28] 孙涛,吴海丰,梁志刚,贺文,张镭,吕平欣,郭秀花.SMOTE算法在不平衡数据中的应用[J].北京生物医学工程,2012,05:528-530.

[29] Dietterich, T.G.Machine-learning research [J].AI magazine, 1997, 18(4): 97.

[20] Polikar, R.Ensemble based systems in decision making[J].Circuits and Systems Magazine, IEEE, 2006, 6(3): 21-45

[31] Breiman L.Bagging predictors[J].Machine Learning,1996,24(2):123-140.

[32] Freund Y.Schapire R E.Experiments with a new boosting algorithm[A] Proceedings of Thirteenth International Conference on Machine Learning[C].Bari:Morgan Kaufmann Publisher,1996:1-9.

[33] Valiant L G.A theory of the learnable.Communications of the ACM,1984,27(11):1134-1142.

[34] 李红莲,王春花,袁保宗.一种改进的支持向量机NN-SVM[J].计算机学报,2003,08:1015-1020.

[35] 郝媛媛,叶强,李一军.基于影评数据的在线评论有用性影响因素研究[J].管理科学学报,2010,08:78-88+96.