学校编码：10384　　分类号_____密级_____
学号：15420141151979　　UDC_____

厦门大学

硕 士 学 位 论 文

惩罚Logistic模型在文本分类中的应用研究

Application of Penalized Logistic Regression Models in Text Classification

黄耀鹏
指导教师姓名：钱争鸣教授
校外导师姓名：孙　宁
专　业　名　称：应用统计
论文提交日期：2017 年 4 月
论文答辩时间：2017 年 4 月
学位授予日期：2017 年 6 月

答辩委员会主席：_____

评阅人：_____

2017 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月

摘要

随着互联网科技的迅猛发展，人类社会所记录的信息呈现“指数级”的增长。海量信息的快速准确分类、查询及个性化推荐等，有着非常迫切的需求。为了更好地解决文本分类任务中的高维稀疏数据分类问题，本文主要做了以下研究工作。

首先，对文本分类各流程所涉及的技术进行了全面的梳理，以准确把握文本分类亟需突破的难点。接着，对惩罚 Logistic 模型进行了理论发展概述；同时，从文献综述的角度探讨了惩罚 Logistic 模型在解决文本分类问题中的可行性，并结合词向量理论和惩罚 Logistic 模型提出一种新的文本分类算法。然后，为了验证惩罚 Logistic 模型在特征选择和分类准确率两方面的能力，将多种惩罚 Logistic 模型与传统特征选择方法、传统分类器进行了文本分类对比实验。并探究了词向量在准确保留中文词语之间的语义信息以及对特征词进行聚类等方面的能力，进而实现了基于词向量分组的惩罚 Logistic 文本分类算法。

从实验分析中，本文主要得到以下研究成果：（1）elastic-net Logistic 等惩罚模型相对于χ^2统计量、F 统计量和互信息等传统特征选择方法，在模型准确率和特征稀疏性上均显示出相当大的优越性；（2）在向量空间模型和词向量两种文本表示框架下，惩罚 Logistic 模型分类效果均与支持向量机相当，并显著优于朴素贝叶斯、决策树等传统分类器；（3）基于词向量分组的 Group LASSO 和 Sparse Group LASSO 模型相比未含特征词先验分组信息的 LASSO 和 elastic-net 模型，模型复杂度大大降低，而准确率却有所提升。

最后，通过网络爬虫抓取中关村在线的手机评论数据，根据文献综述研究与实验结论挑选适当的模型，从情感分类和评论有用性的影响因素两个角度进行了手机评论挖掘的应用研究。最终本文为实现一个具有商业应用价值的在线商品评论挖掘系统提供了较全面的模型支持。

本文的贡献主要有：一方面，将惩罚 Logistic 模型引入文本分类领域，并创新性地提出一种能够结合词向量理论与惩罚 Logistic 模型优点的文本分类算法，拓宽了惩罚 Logistic 模型的应用领域；另一方面，从多个角度进行了手机评论挖掘的应用研究，该部分研究成果可以为在线商品平台建立评论推荐系统提供参考价值。由于资源和时间的限制，本研究还存在以下不足之处：基于词向量分组的惩罚 Logistic 文本分类算法还需要在运行效率上进行优化；有用性影响因素模型可以进一步考虑交叉效应等方面的影响。

关键词：文本分类；惩罚 Logistic 模型；产品评论挖掘

Abstract

With the rapid development of Internet technology, the information in human society increases exponentially. Fast and accurately classification and recommendation of text information are in great demand. In order to solve the high dimension and sparse data problem faced by the text classification task, our paper mainly do the following research work.

Firstly, we carry out a comprehensive introduction of technologies involved in the text classification process. Secondly, we summarize the theoretical development of the penalized logistic regression model, and discuss the feasibility of the penalized logistic regression model in solving the text classification problem from the perspective of literature review. In addition, we propose a new algorithm to combine the word vector theory and the penalized logistic regression model for text classification. Thirdly, we make comparisons of the penalized logistic regression models with traditional feature selection methods and traditional classifiers on experiment, and realizes the text classification algorithm we proposed.

In the experimental analysis, the following results are obtained. (1) Compared with the traditional feature selection methods such as χ^2 statistics, F-statistics and mutual information, the elastic-net logistic is more superior at both accuracy and sparsity. (2) The penalized logistic regression models are comparable with the support vector machine at classification performance, and outperform other traditional classifiers such as Naive Bayesian and decision tree under the framework of both vector space model and word vector. (3) Comparing with the LASSO and the elastic-net models, the model complexity of the Group LASSO and Sparse Group LASSO models are greatly reduced and the classification accuracy are improved.

Last but not least, we crawl the mobile phone review data from *Zhongguancun Online* through Web crawler technology, and make an empirical study on online product reviews mining. First, we establish a sentiment analysis of product reviews based on text classification models. Second, we make a comprehensive study of extracting predictors for product review helpfulness, and build a quantitative model to measure the helpfulness of product reviews. Based on our models, we can classify and rank the product review scientifically.

Keywords: Text Classification; Penalized Logistic Regression; Product Review Mining.

目录

Contents

第一章　绪论

1.1 研究背景

随着互联网科技的迅猛发展，人类社会所产生的信息呈现爆炸式增长。在我国政府主导的“互联网+”产业升级背景下，海量信息的快速准确分类、查询及个性化推荐等，显得至关重要。从个人角度来说，信息的快速准确查询，有利于日常生活的便利化；从企业角度来说，及时准确地掌握有效的信息，有利于提升企业的生产效率；从国家的角度来说，信息的准确分类和及时有效的传播关系到社会的稳定和发展。

海量文本信息的检索和挖掘利用，所面临的一大基础性问题是文本的分类问题。文本分类问题指的是通过统计学或机器学习的方法，将一个文档准确的归类到所属的类标签中，它属于一种有监督学习的问题。现今，文本分类在文本管理、搜索引擎信息检索、社会舆情分析、电子商务平台的产品评论挖掘、垃圾邮箱的过滤、以及语义挖掘等领域均扮演者重要角色[1][2]。在互联网时代来临之前，人们主要依靠人工处理来对信息进行归类汇总和提取。而随着当今社会的信息化渗透到每一个角落，人们逐渐被淹没在海量的信息洪流之中。在智能手机大量普及的移动互联网时代，信息增长更趋猛烈、垃圾信息泛滥成灾。而在不久的将来，物联网的技术更会将信息社会推向一个新的高峰，信息增长的速度将会更加超越我们的想象。在这样的背景下，对信息分类处理智能化的技术要求也不断提高，依靠人工分类的低效率方法已经远远无法满足社会对高效获取有用信息的需求。

目前，文本自动分类技术的商业前景非常宽广。近几年来，我国商业领域已涌现出一批依赖于文本自动分类技术的互联网产品。如 2012 年 8 月上线的“今日头条”应用就是一个典型的代表。它是一款基于数据挖掘技术的新闻推荐产品，其主要特色功能是通过对成千上万的网站进行文本挖掘与分类，智能化地给用户推荐高质量的新闻资讯。又如 2011 年 1 月创立的社会化问答网站“知乎”，对本

文自动分类技术也有着非常深入的应用。“知乎”的各大基于推荐引擎的产品如“知乎日报”、“发现”栏目，以及每天的优质内容推送等，都离不开文本自动分类技术的应用。

随着网络购物的普及化，我国网络购物的用户数量于 2015 年 12 月达到了 4.13 亿，相比 2014 年增长了 14.3%。据调查显示，43.8%的网民表示喜欢在互联网上发表评论信息[49]。据笔者观察，在中国最大的 3C 数码产品购物网站京东商城上，一款时髦的手机产品 Apple iPhone 6s 的产品评论高达 57.6 万以上。然而，大量的商品评论的存在同时也会带来一些负面影响。由于购物评价是一个相对自由而无约束的行为，这样会导致一部分购物者随意评价、商家刷评等现象的发生，使得购物网站充斥大量的垃圾评论和重复评论。如何帮助消费者在泥沙俱下的众多评论信息中迅速查阅到客观翔实的高价值产品评论，是一个亟需解决的难题。

1.2 研究意义

由于文本分类问题具有语义复杂、数据非结构化、超高维度、高度稀疏化等特点，如何构建一个能像人类大脑一样理清语义、并且能够有效处理非结构化高维稀疏数据的文本分类模型仍是我们当今社会面临的一大难题。

总的来说，文本分类问题可以归结为以下几大步骤：（1）分词（Word Segment）；（2）文本形式化（Representation）；（3）特征抽取（Feature Extraction）及特征选择（Feature Selection）；（4）分类器的构造（Machine Learning）；（5）分类效果的评价（Performance Evaluation）。文本分类流程如下图所示。

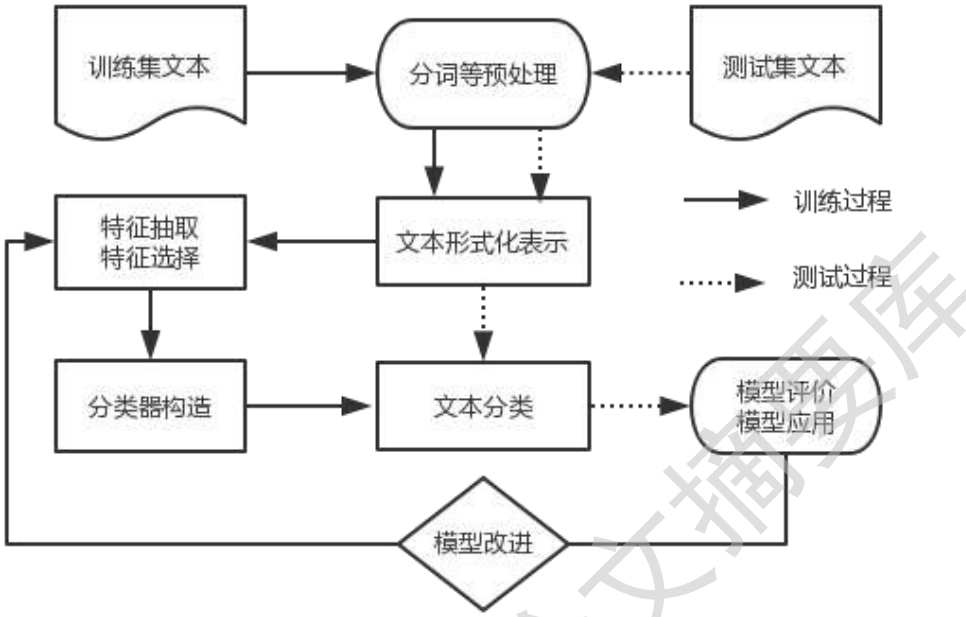

图 1-1：文本分类流程图

随着相关研究技术的发展，分词方法已经取得较为显著的成果；文本形式化表示技术在最近几年也有重大突破；而特征选择和分类器的构造上，一直进展缓慢，各种方法的优劣没有统一的定论。另一方面，近十几年来，统计学领域在处理超高维数据分析的问题上，取得了长足的进步，积累了丰富的方法和经验。因此，在方法创新上，我们可以借助统计学领域发展成熟的变量选择方法，与文本分类领域的前沿理论相结合，构造出更加智能化的文本分类算法。

文本分类技术在电子商务平台产品评论挖掘中的应用已经逐渐开始被重视起来。美国最大的电子商务平台之一亚马逊网站已经实现了简单的商品评论排序和推荐功能。然而，国内的电子商务平台对评论的挖掘和推荐还不够重视，还没有形成成熟的评论过滤、排序和推荐功能。我们可以借助文本分类算法和统计分析技术，实现对海量评论数据的有效过滤和排序。这样，会使得本文的研究同时具有较大的学术研究意义和潜在的商业价值。

1.3 研究内容

1.3.1 研究框架

本文的研究框架可见图 1-2，各章节的内容概述如下。

第一章介绍研究背景，并阐明本文的研究目的及意义；与此同时，简要概括论文的整体研究框架、主要研究内容及研究特色。

第二章从中文分词、文本形式化表示、文本特征降维、分类器的构造等方面进行发展进程研究和理论概述。

第三章对惩罚 Logistic 模型进行理论梳理，并提出一种能够结合词向量理论和惩罚 Logistic 模型的文本分类算法。

第四章从多个角度设计文本分类实验，分析惩罚 Logistic 模型在文本分类任务中的优越性。

第五章将文本分类技术应用于在线商品评论挖掘。首先针对在线商品评论数据建立情感分类模型，实现评论的情感倾向判断；接着，对评论质量的影响因素进行全面综合的统计分析，建立评论有效性影响因素量化模型，从而实现对评论质量的量化和合理排序。

第七章对论文的整体内容进行全面的总结和展望。