2018 届硕士专业学位研究生论文（全日制研究生）

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：51151201070

華東師範大學

East China Normal University

硕士专业学位论文

Master's Degress Thesis (Professional)

论文题目：基于众包的主动学习模型优化方法及应用

院　　系：计算机科学与软件工程学院

专业学位类别：计算机技术

专业学位领域：大数据分析与知识处理

论文指导教师：杨静 副教授

论 文 作 者：陈博闻

2017 年 9 月完成

Dissertation for Master Degree, 2018 School Code: 10269

Student ID: 51151201070

EAST CHINA NORMAL UNIVERSITY

Optimization Method and Application of Active Learning Model Based on Crowdsourcing

Department: School of Computer Science and Software Engineering

Major: Computer Technology

Research Interest: Big Data Analysis and Knowledge Processing

Advisor: Associate Prof. Jing Yang

Mater Candidate: Bowen Chen

Fulfilled in September 2017

华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于众包的主动学习模型优化方法及应用》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：_____ **日期：** 年 月 日

华东师范大学学位论文著作权使用声明

《基于众包的主动学习模型优化方法及应用》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的著作权归本人所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和学校指定的相关机构送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校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（请勾选）

（ ）1.经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文[1]，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。

导师签名_____ 本人签名_____

年 月 日

[1] “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门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）。

陈博闻 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
顾君忠	教授	华东师范大学计算机科学与软件工程学院	
胡琴敏	副教授	华东师范大学计算机科学与软件工程学院	
崔修涛	高级工程师	中电科软件信息服务有限公司	

摘 要

随着 Internet 的飞速发展，海量的互联网数据以爆炸式的速率迅速膨胀，数据处理、人工智能、机器学习等技术的不断进步使得计算机处理这些信息的能力日益增强。然而，单纯依赖机器算法还很难完全达到人类的认知水平，而众包技术能够借助群体智慧，利用人类优势来解决机器难处理的问题，例如进行文本情感分析、图像标注等。同时，我们希望人类的智慧能不断地“指导”机器进步，基于众包的主动学习模型就是这个思想的一个典型代表。基于众包的主动学习模型将机器学习算法和人工知识补充结合在一个算法流程中，它的主要目的是在标注成本有限的情况下尽可能获得更优的模型。

为了在一次众包标注任务中获取人类更多的潜意识信息用于提升机器的学习能力，研究人员提出了多种众包任务收集不同形式的反馈信息。然而，这些方法有的过于复杂难以构建，有的需要用户创造性地给出潜意识的知识导致用户体验差而任务难以顺利结束。因此，本文从任务设计角度出发，在任务执行过程中通过尽可能简单的操作，收集用户的解释性反馈信息从而获取更多人类的知识，并且将其融入分类模型中提高模型的学习性能。

另一方面，由于众包工作者的不确定性，我们需要进行众包质量控制，现有的优化工作均局限于众包阶段，仅对众包过程中的标签质量或人群质量进行控制，这些方法将所有待标注数据集未做筛选地交由众包进行标注没有将众包和主动学习很好地结合。因此，我们提出一种跨阶段的优化方法，考虑了主动学习模型中的众包任务和传统众包任务的区别，将众包与主动学习更好地结合。该方法利用众包知识改进现有的主动学习采样策略，在保证样本具有充足信息量和代表性的基础上，尽可能挑选适合大众标注的样本，减少众包标签噪音，提高标签质量，从而改善模型预测效果。

本文主要贡献如下：

1 提出一种基于众包解释性反馈的主动学习模型优化方法，该方法设计一种交互良好的众包任务，任务过程中收集规则化的众包解释性信息，通过解释性信

息挖掘人类对数据的潜在认识，并且将这些反馈信息融入到分类模型中以提升模型学习能力。

2 为了将众包技术更好地融入主动学习分类模型，我们对主动学习挑选策略进行改进，引入众包标注置信度，挑选更适合人群的样例进行标注，从而减少众包标签噪音，达到提高分类模型性能的目的。

3 将本文的方法运用于两个文本分类数据集上，通过多组对照实验，验证该方法的有效性，并对实验结果加以分析，提出未来的改进工作。

关键词：众包、解释性反馈、主动学习、众包置信度、文本分类

ABSTRACT

With the rapid development of the Internet, massive Internet data expand at explosive rate expansion. The unceasing progress of data processing, artificial intelligence, machine learning and other technologies enhance the ability of the computer to progressing information. However, relying solely on the machine algorithm is still very difficult to fully meet the level of human awareness. Crowdsourcing technology can use group wisdom and human advantages to solve the problems which computers can not figure out well, such as text emotion analysis, image labeling and so on. At the same time, we hope that human wisdom can "guide" the machine constantly, active learning model based on crowdsourcing can achieve this idea. Active learning model based on crowdsourcing combines the machine learning algorithm with the artificial knowledge in an algorithmic flow. Its main purpose is to get a better model as much as possible when the cost of labeling is limited.

In order to obtain more human subconscious information in a crowd task to improve the performance of the machine learning, the researchers proposed a variety of crowdsourcing tasks to collect different forms of feedback. However, some of these methods are too complex to build, and others fails due to lack of the subconscious knowledge lead provided by users in a creative way. Therefore, this paper, from the perspective of the task design, collects the user's explanatory feedback in the process of task execution to obtain more human knowledge through the operation as simple as possible, and integrate it into the classification model to improve the learning performance of the model.

On the other hand, due to the uncertainty of crowdsourcing workers, we need to carry out mass quality control of crowdsourcing. The existing optimization work only imposes restrictions on the crowdsourcing stage, such as the quality of labeling and the crowd. Those methods marked all the data without filtering and did not combine the

crowdsourcing and active learning well. Therefore, we propose a cross-stage optimization method, consider the differences between crowdsourcing tasks in the active learning model and the traditional task of the package, and better integrate the crowdsourcing and active learning. This method makes use of the knowledge from crowdsourcing to improve the existing active learning sampling strategies. On the basis of ensuring the sufficient information and representation of the samples, it can select the samples which are as suitable as possible for the public label, reduce the noise of the crowd label and improve the quality of the crowd labeling, so as to improve performance of model prediction.

The main contributions of this paper are as follows:

1. This paper proposes an optimization method of active learning model based on crowd's explanatory feedback. We design an well-interactive method of crowdsourcing tasks, collecting the rule of explanatory information in the task process, and explore the human subconscious on data in order to improve the model performance.
2. In order to better integrate crowdsourcing into the active learning classification model, we improve the query strategies by importing the crowd confidence, selecting more suitable samples for the crowd to label, so as to reduce the noise of crowdsourcing labeling, and improve the model performance.
3. The method of this paper are used in two data sets for text categorization, and verifies the effectiveness of the method by multiple sets of control experiments. We analyzed the experimental results and discuss the future works.

Keywords: Crowdsourcing, Explanatory Feedback, Active Learning, Crowd Confidence, Text Classification, Data Crowdsourcing

目 录

摘 要 I

ABSTRACT III

第一章 绪论 1

1.1 研究背景与意义 1

1.2 研究现状及挑战 2

1.3 本文的主要工作 3

1.4 本文的组织结构 3

第二章 相关研究工作 5

2.1 众包技术 5

2.1.1 众包概念 5

2.1.2 众包平台 5

2.1.3 众包反馈形式 7

2.2 主动学习 8

2.2.1 主动学习概念 8

2.2.2 主动学习的样例选择算法 10

2.2.3 样例选择优化方法的讨论 12

2.3 基于众包的主动学习模型 13

2.4 本章小结 14

第三章 基于众包解释性反馈的主动学习模型优化方法 15

3.1 方法的引出 15

3.2 概念定义及方法流程框架 16

3.2.1 众包反馈与解释性反馈 16

3.2.2 方法流程框架 18

3.3 众包解释性反馈的收集及主动学习模型改进方案 19

3.3.1 众包解释性反馈的收集 19

3.3.2 融入众包解释性反馈的主动学习模型 20

3.4 本章小结.....23
第四章 融入众包标注置信度的主动学习算法改进方法.....24
4.1 方法的引出.....24
4.2 概念定义及方法流程框架.....25
4.2.1 众包偏差与众包标注置信度.....25
4.2.2 方法流程框架.....25
4.3 融入众包标注置信度的采样策略改进方法.....27
4.3.1QUIRE 算法介绍.....27
4.3.2 融入众包标注置信度的采样策略——QUIRE_CrowdDiff.....27
4.4 本章小结.....30
第五章 实验设计与结果分析.....31
5.1 众包过程.....31
5.2 评价指标.....33
5.3 实验设计.....34
5.4 基于解释性反馈的主动学习模型优化方法实验.....37
5.4.1 不同主动学习算法下的讨论.....37
5.4.2 不同分类器和不同标签分布下的讨论.....38
5.4.3 不同权重系数的讨论.....39
5.4.4 迭代式环境下比较不同方法的分类效果.....40
5.4.5 不同 batch size k 对模型性能的影响.....43
5.4.6 标注成本和准确率的讨论.....44
5.5 融入众包标注置信度的采样策略改进方法实验.....45
5.5.1 不同挑选策略下的讨论.....45
5.5.2 不同挑选策略下众包标注情况.....46
5.5.3 不同影响因子下准确率的变化情况.....46
5.6 本章小结.....47
第六章 总结和展望.....48

6.1 本文工作总结.....48
6.2 未来工作.....48
参考文献 50
附录一 作者攻读硕士学位期间参与的科研项目与专利 56
致谢 57

第一章 绪论

1.1 研究背景与意义

近年来，越来越多的人开始意识到数据对科技、企业的重要性。然而，爆炸式的数据增长让人们陷入了“数据丰富、知识匮乏”的窘境。虽然人们逐渐学会利用机器学习中的模型来挖掘数据背后的知识，然而，训练高性能的模型仍然依赖于大量有标签的样本集。但是在实际情况下，获取没有标签的样本集相当容易，而获取大量的有标注的样本集却费时费力。为了获取高质量带标签的样本通常是雇佣领域专家手工进行标注，但是这种方式需要支付较高的佣金而且任务周期较长。

众包技术的出现[17]，让低成本高效率地完成机器难处理的任务成为可能。“众包（crowdsourcing）”是一种公开面向互联网大众的分布式的问题处理机制[51]，它允许数据需求者作为任务发布者将数据任务发布到互联网上，利用在线大众的智慧来解决问题，此时任务发布者只需要支付少量的报酬即可获得想要的数据标签，相比较雇佣专家的方式能够节省大量的成本和时间开销。此外，众包面向的工作对象是人，所以它更适合解决机器难处理且专业性并不特别强的任务，例如情感极性分类任务，实体匹配任务以及一些简单的图像标注任务。

虽然引入了众包技术，但若将所有的标注请求都交给众包进行标注同样需要花费大量成本。于是，研究人员们[1,5,9,12,15,16,26,27,32]提出了主动学习模型，挑选一部分较优的样本交给众包平台标注，并设计合理的任务，获取更多大众潜在知识，使得在保证模型性能的同时减少标注成本。基于众包的主动学习技术允许人们在少量的标注集的限制条件下也能训练出较好的模型，它具有很大的商业价值和应用前景。

本文主要针对文本分类问题，分别从众包任务设计和主动学习挑选策略角度出发，研究基于众包的主动学习文本分类模型的优化方法，目的是在挖掘更多用户潜在认识的同时，将众包技术更好地融入主动学习分类模型中，从而降低标注

成本并提升模型性能。

1.2 研究现状及挑战

众包研究领域中，为了挖掘更多来自人类的潜意识知识，相关人员设计了多种众包任务，收集不同形式的反馈信息，并从中挖掘有用信息融入到模型中。这些方法[7,41,20,14,8]主要通过增加任务中的用户交互行为扩大众包反馈内容，从而获取更多的人类知识。这些方法中的反馈形式主要包括以下两种：内容反馈（文本或标签）和行为反馈（点击或拖拽）。文献[7,41]通过组合不同问题类型来设计众包任务，但这些任务有的在设计过程中没有很好地考虑用户体验，有的方法仅适合特定数据集。文献[20,14,8]让用户以玩游戏闯关的方法，挖掘隐藏在用户行为背后的潜意识知识。然而，这些众包任务的设计均较复杂而且实现成本较大。因此，本文希望通过一种交互良好且易于实现的任务形式来扩充众包反馈的内容使得众包收集的内容不再局限于标签。文本设计规则化的形式让众包工作者给出解释性信息，从而挖掘人类对数据的潜在认识，并且将这些反馈信息融入学习模型中，充分利用众包反馈的内容实现提高分类器性能的目的。

当面临训练集不够且标注成本有限的问题时，为了以低成本的方式得到尽可能优的学习模型，人们已经提出了利用众包和主动学习相结合的方式解决问题。目前的众包与主动学习相结合的优化工作并不多[29,47,50,30,39]，且这些方法或仅专注于优化主动学习挑选策略，或仅专注于优化众包标签质量，这些方法都是从两个阶段分别进行优化，并没有将众包和主动学习模型很好地结合起来。为了将众包技术更好地融入主动学习分类模型中，本文对主动学习挑选策略进行改进，引入众包标注置信度这一衡量指标，从而减少众包标签噪音以达到提高分类模型性能的目的。

本文对基于众包解释性反馈的主动学习文本分类优化方法的研究工作主要面临如下两个挑战：

1. 人们常常通过潜意识来判断一些主观意识问题。如何准确并有效地激发人们表达对文本的潜在认识是相当困难的。另外，复杂化的交互操作会使得用户

体验大打折扣以至于影响任务完成的效率和质量。因此，如何设计一种交互良好的众包任务挖掘更多用户潜意识知识是值得思考的问题。

3. 如何将众包技术更好地融入到主动学习分类模型中，挑选出更适合大众的样本从而减少众包标签噪音，改善模型性能。

1.3 本文的主要工作

本文研究目标是提高文本分类器的效率和性能，主要研究内容包括：

1. 从众包任务设计角度出发，挖掘更多人类对文本数据的潜在认识；

2. 从主动学习采样策略角度出发，挑选更适合人类的标注样本，改善众包标签质量。

针对以上研究内容，本文主要工作如下所示：

1. 提出一个基于众包解释性反馈的主动学习文本分类模型优化策略。该方法设计一种交互良好的众包任务，充分利用众包交互性，收集规则化的解释性反馈，挖掘众包工作者对文本数据的潜在认识。

2. 将众包技术和主动学习模型更好地结合，在样例选择过程中引入众包标注置信度这一新的衡量标准，在确保挑选出信息量大代表性强的样本的同时，挑选适合大众标注的样本，降低来自众包的标签的错误率，从而优化模型输入。

3. 将本文的方法运用于两个文本分类领域数据集上，通过多组对照实验，验证该方法的有效性，并对实验结果加以分析，提出未来的改进工作。

1.4 本文的组织结构

本文组织结构如下：

第一章论述了本文的研究背景与意义和研究现状及挑战，介绍了本文的主要工作，并对本文的整体结构进行总结。

第二章主要对本文研究过程中涉及到的相关研究工作进行综述，总结并分析了现有方法的缺陷。

第三章首先通过对文本数据进行分析引出本文基于众包解释性反馈的主动

学习优化方法，然后详细介绍了该方法中的相关概念，并给出了算法框架。

第四章首先对真实实验结果进行分析引出融入众包标注置信度的主动学习算法的主要思想，并详细介绍了方法中的相关概念和算法框架。

第五章将本文提出的方法应用于两个文本分类数据集上，通过多组对照实验验证其有效性和可行性。

第六章总结了文本的研究成果，讨论了本文研究的不足之处并对未来工作进行展望。

第二章 相关研究工作

2.1 众包技术

2.1.1 众包概念

“众包”一词由美国连线杂志的记者 Jeff Howe 于 2006 年首次提出，他对“众包”的定义为：“一个公司或机构把过去由员工执行的工作任务，以自由自愿的形式外包给非特定的（而且通常是大型的）大众网络的做法。众包的任务通常由个人来承担，但如果涉及到需要多人协作完成的任务，也有可能以依靠开源的个体生产的形式出现。”[17,56]众包这一互联网概念提出了一种新的企业模式，是一种基于互联网大众的分布式问题解决机制[17]，它通过将互联网上人群提供的未知数据集成在一起，完成计算机难以处理的任务。事实上，现实生活中存在着大量的依赖人的主观意识但计算机难以处理的问题，例如语言情感标注[33,21]、图像标注[44]、实体匹配[3,6,28,34,45]等。该类问题可以通过众包将任务发布到互联网上，让互联网上未知的大众来合力解决。借助众包技术我们可以大大提高这些任务的处理效率和结果质量。因此，众包技术有着广大的应用前景和市场。在近年来，众包技术受到越来越多企业的青睐，也日益成为相关研究人员的研究热点。

2.1.2 众包平台

互联网人群参与众包的形式主要有两种：协作式众包（collaborative crowdsourcing）和竞赛式众包（crowdsourcing contest）[51,49]。第一种形式中任务需要大众协作完成，并且没有奖励回报，其中较为著名的案例有：维基百科 Wikipedia[2]和 reCAPTCHA[3]。维基百科是一个开放的、自由的网络百科全书，允

[2] https://zh.wikipedia.org/wiki/Wikipedia

[3] https://zh.wikipedia.org/zh/ReCAPTCHA

许大众进行在线编辑与修改。reCAPTCHA 项目是由 CMU 教授 von Ahn 提出的[51]，通过在验证码中嵌入书籍的扫描信息来完成纸质书籍的电子化。而竞赛式众包任务通常由一个人独立完成，并且在完成任务后能获取一定金钱报酬，其典型的案例有 AMT[4]（Amazon Mechanical Turk）和 CrowdFlower[5]。在竞赛式众包模式下，任务发布者被允许在众包平台上发布任务与需求，任务工作者可以在该平台上挑选任务并执行，任务完成后可以获得相应的报酬。AMT 和 CrowdFlower 是目前主流的众包平台，其常见的任务有：数据标注、数据收集等。

众包的工作流程由任务发布者、工作者和众包平台共同完成，具体流程如图 2.1 所示。

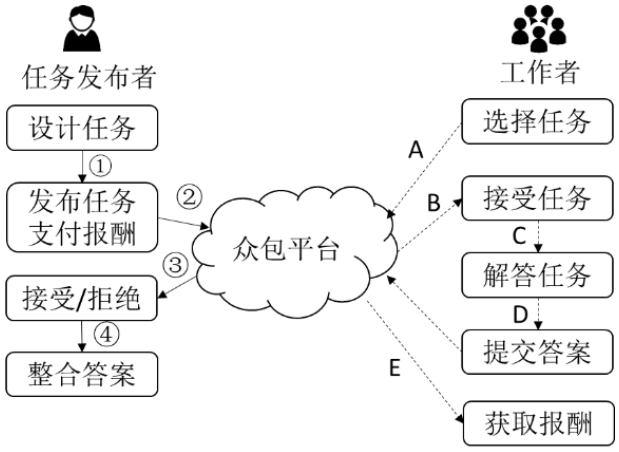


图 2.1 众包的工作流程

如图 2.1 所示，当任务发布者想要利用众包来实现自己的需求，需要执行如下步骤：

① 设计众包任务；

② 发布众包任务并支付一定报酬；

③ 拒绝或接受工作者的答案；

④ 按照自己的需求整合答案，完成任务。

当工作者想要接受众包平台上的任务，并获取相应报酬时，需要执行以下步骤：

[4] https://www.mturk.com/mturk/welcome

[5] https://www.crowdflower.com/

A 通过众包平台选择感兴趣的任务；

B 接受任务；

C 解答任务；

D 提交任务答案；

E 答案被接收后获取应有报酬。

众包平台主要目的是执行任务和收集答案。目前主流的商用众包平台包括Amazon Mechanical Turk（AMT）、CrowdFlower、CloudCrowd[6]等。国内近年来也涌现了很多众包平台，例如脑力库、猪八戒[7]、阿里众包[8]、百度众包[9]等。商用众包平台会针对任务发布者和工作者两种不同的身份人群的不同需求提供相应的服务。任务发布者可以通过众包平台发布任务并支付一定的费用，工作者可以在平台上选取任务来完成并获得一定的收益。在这些众包平台上，发布的任务主要是一些粒度较小的微观任务，因为工作者们往往更青睐于利用他们的碎片化时间去完成一些简单的，粒度小的任务，例如数据标注任务。本文实验中的众包任务，将借助阿里众包平台，发布文本分类数据的标注任务。

2.1.3 众包反馈形式

引入众包的目的是利用人的先天优势来解决机器难处理问题，为了挖掘更多人类潜意识的知识来提升机器的学习能力，许多研究人员开始从众包任务设计角度来获取更多的人类信息。他们均通过增加用户交互行为来扩大众包反馈内容，从而获取更多的人类知识。增加的交互行为主要包括文本或标签的内容反馈以及点击或拖拽的行为反馈两种形式。

文献[7,41]中提出一种组合的任务模式，其中每个任务包含多个问题，如填空题、选择题等。让用户提供多种形式的内容反馈，包括文本、标签等。但是文献[7]中的众包任务仅适合于图像数据集，而文献[41]中需要用户填写文本类型数

[6] http://www.cloudcrowd.com/

[7] http://www.zbj.com/

[8] http://zhongbao.alibaba.com/

[9] http://zhongbao.baidu.com/

据，该方式增加了用户的行为代价从而极大的降低了平台的用户体验。

还有一些方法如文献[20,14,8]，通过设计游戏的方式去挖掘人类的潜在认识。文献[20]设计了一款“气泡”游戏，该游戏给用户提供一个仅能看清轮廓的模糊图像，用户需要有偿的揭示模糊图像中的“气泡”大小的区域，然后做出分类决定。这种任务形式不仅能够从人类行为中获取有辨识度的特征并将其运用于机器分类中，还能够一定程度减少用户交互从而降低标注成本。然而这种形式仅适用于图像数据集，不适用于文本数据集。文献[14]中建立一种自动生成 3D 游戏场景（如厨房、卧室）的模型，允许用户将相关物体对模型进行填充，该方法通过众包游戏收集到的物体之间的相对空间位置关系，从而形式一个符合常识性知识的游戏场景。文献[8]同样利用 3D 游戏获取游戏编辑者和玩家的潜意识中的常识性知识。但是，后面两种方法都需要设计一套完整的游戏，构建所需代价太大且收集的反馈信息太多太杂导致后期难以处理。

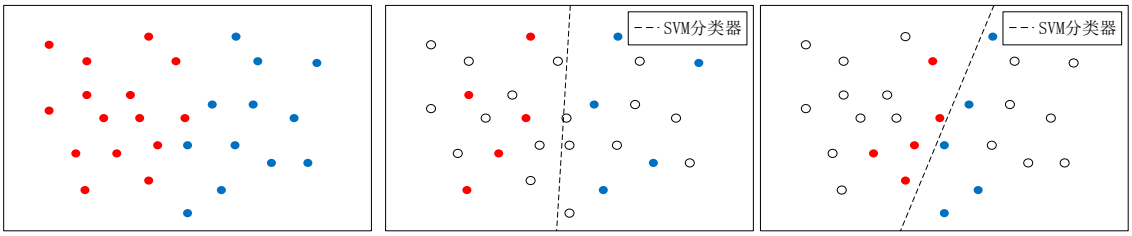
对不同形式反馈的众包任务的分析中，我们发现在获取更多用户反馈的同时需要考虑任务搭建的难度、用户体验度以及反馈信息的处理难度等多种因素。因此，本文提出一种交互良好且易于实现的众包任务形式，该方法通过收集多种内容反馈（包含标签反馈和解释性反馈），去挖掘更多人类潜意识的知识，从而提升机器的学习能力。方法的具体内容我们将在第三章详细说明。

2.2 主动学习

2.2.1 主动学习概念

主动学习[52]是机器学习领域的一个重要分支。在机器学习领域中，分类模型的效果依赖于标注数据集，然而，要获取大量的有标签数据集需要花费大量的人力和时间成本。为了减少标注样本集、降低标注成本，研究人员引入了主动学习算法优化模型。图 2.2 所示一个二分类问题，从图中我们可以看出，不同的训练集对模型的学习性能影响较大，其中主动学习挑选并标记的样本比随机挑选的样本能够得到精度更高的分类器。由此可见，主动学习能够挑选出更优的样本优化模型，并一定程度地降低标注数。主动学习的主要思想是在模型训练的过程中

主动提出一些“标注请求”，发现一些能够最大程度提升模型性能的优质样本进行标注并交由模型学习，反复迭代直到标注代价或模型精度达到一定标准为止。主动学习流程如图 2.3 所示。与传统的监督方法相比，主动学习能够在训练过程中挑选出有辨识能力的样本点，更好地处理较大数据集，并且减少训练数据量，降低人工标注成本。



(a) 某二维数据集类别真实分布情况　(b) 随机挑选10个点作为训练集训练而得的分类模型　(c) 利用主动学习算法挑选10个点作为训练集训练而得的分类模型

图 2.2 主动学习示例

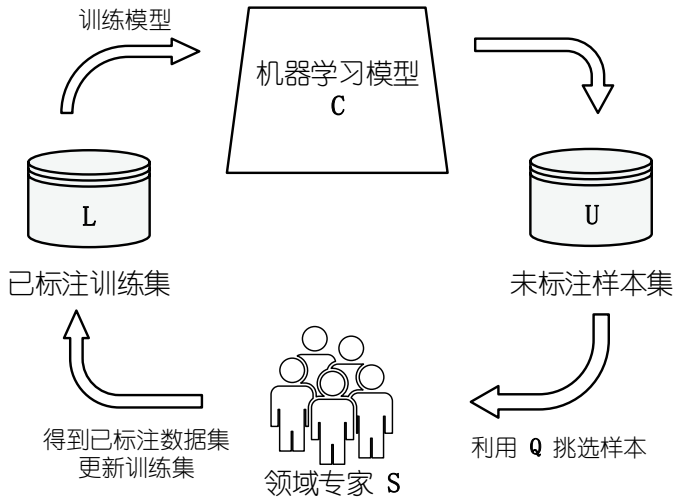


图 2.3 主动学习流程图

主动学习算法由以下五个组件进行建模：

$$A=(C,L,S,Q,U)$$

其中 C 是一个或一组分类器；L 是已标注训练集；S 是监督者，即对未标注样本进行标注的人；Q 是采样策略，用于在未标注样本中挑选信息量大的样本；U 是剩余未标注样本集。[52]

主动学习算法过程分为两个阶段：

第一个阶段为初始化阶段，该阶段的目的是基于原始已标注训练集样本训练

初始分类器模型；第二个阶段为循环样例选择阶段，从未标注样本集 U 中，根据某种采样策略 Q，选取一些未标注样本进行标注并加到训练样本集 L 中，重新训练分类器，直到达到训练停止标准为止。

主动学习算法是一个迭代的过程，分类器使用迭代时反馈的样本进行训练，不断提升分类效率。

2.2.2 主动学习的样例选择算法

主动学习的样例选择算法主要分为以下三种：成员查询综合算法、基于流的样例选择算法和基于池的样例选择算法。

成员查询综合算法，由文献[2]中首次提出，允许学习器挑选输入空间中的任意未标注样本进行标注，该算法的缺点在于挑选样本时没有考虑样本的分布情况。此后，相关研究人员又提出了基于流的样例选择算法[4,9]，即每次对一个新样本进行评估判定是否进行标注，该算法的缺点是过程中往往需要设置衡量信息量的阈值，缺乏普适性。基于以上方法的缺点，研究人员提出了基于池的样例选择算法[10]，该算法的主要思想是构建一个待标注样例池，通过评估池中所有样本，挑选最优的一个或多个样本进行标注，其中常用的评估策略是挑选信息量大的样本进行标注。基于池的样例选择方法从一个样例池中挑选最优样本，克服了前两种方法的不足，因此在文本分类[10,1,43,15]、信息检索[40,36]、图像分类检索[42,48]、癌症诊断[27]等领域中均受到了广泛的运用[37,54]。

基于池的采样策略方法，根据挑选标准不同分为三种：基于不确定性的采样策略[10,23]，基于版本空间缩减的采样策略[38]以及基于最小期望误差的采样策略[35]。其中，基于不确定性的采样策略是适用性最广的一类采样策略，本文将对基于不确定性的采样策略进行研究，并提出改进方案。

基于不确定度的样例选择是最简单最常见的一种挑选样本的策略。顾名思义，这种样例选择算法的主要思想是挑选对模型而言分类最不确定的样本，将这些样本交由专家进行标注后，将收集到的标签加入训练集重新训练一个新的模型。该采样策略适用于不同类型的分类器。不同的分类模型，挑选样本的决策函数会演

化成不同形式，下面我们介绍不同分类器下的挑选策略。

当运用概率模型处理一个分类问题时，模型会计算出待分类样本的后验概率，这反映了预测类别的确信度，基于不确定度的样例选择算法会挑选出确信度低的样本拿去标注。对于一个包含三个或三个以上类别的问题，一个通用的挑选策略由公式 2-1 表示：

$$x_{LC} = \arg\max_{x} 1 - P_\theta(y^* \mid x) \tag{2-1}$$

其中 $y^* = \arg\max_{y} P_\theta(y \mid x)$ ，即为分类模型对样本预测的标签。

然而以上方法仅仅考虑了样本可能性最大的类别，当两个最大的类别概率值相差非常小时，同样说明了模型对该样本预测的不确定度相当高，因此该策略不适合处理多分类问题。直观地看，如果一个样本处于两个样本的边缘，那么它对于机器而言预测结果会是模棱两可的，如果此类样本被正确地标注并用于模型训练中，那么模型的分类效率必定会大大提升。利用形如公式 2-2 的方法来表示这种边缘样例选择策略：

$$x_{BT} = \arg\min_{x} P_\theta(y_1 \mid x) - P_\theta(y_2 \mid x) \tag{2-2}$$

其中 y_1, y_2 分别是分类模型预测出的概率最大的两个类别。上述方法修正了第一种方法的缺陷，将第二可能的类别概率也考虑其中。然而，对于大规模的数据集而言，边缘选择的方法往往会忽略数量类别的分布。

一个更常用的不确定度样例选择策略引入了信息熵来反应样本的不确定度，计算公式如 2-3 式：

$$x_H = \arg\max_{x} -\sum_{i} P_\theta(y_i \mid x) \log P_\theta(y_i \mid x) \tag{2-3}$$

这里的 y_i 表示 x 所有可能的分类。信息熵是信息论中对信息的量化，这种思想经常运用于机器学习中对样本不确定性的度量。

基于不确定度的样例选择算法同样也适用于非概率分类模型。在支持向量机模型（support vector machine，SVM）中，通过一个超平面最大化分隔样例空间实现对数据进行分类，其主动学习策略是：利用边缘抽样（Margin Sampling）[43]的方法根据数据点到超平面的距离来衡量该样本的不确定度，然后挑选离超平面最近的样本进行标注。直观地看，距离划分平面越近就意味着数据样本的不确定度越大。如图 2.4 所示(实心点为无标签的数据点、实线为训练得到的分类模型)，因此收集越靠近超平面的样本标签对模型的性能提升更大。许多主动学习相关文献中都用支持向量机模型作为其基础训练模型。[43,16,26,22,25]

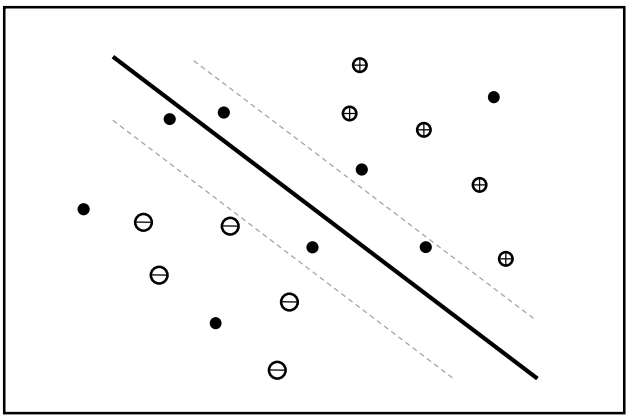


图 2.4 SVM 分类模型图

假设目前面临的是二分类问题，基于边缘采样的主动学习算法挑选待标注样本的公式如 2-4 式所示：

$$x_{MS}^{*} = \arg\min_{x_i \in D_u} \sum_{j=1}^{l} a_j y_j K(x_j, x_i) + b \tag{2-4}$$

以上边缘抽样算法通常能够充分利用学习模型的特性，简明高效且便于理解，因此被广泛使用。然而这种算法也存在一定的缺陷，在迭代初期模型质量低的情况下，挑选过程往往会偏向于挑选具有特定性质的样本，在局限于特定数据集上的预测准确率会大幅降低模型性能。另外，不符合原始分布的异常点（outlier）同样会对模型造成较大的负面影响。

2.2.3 样例选择优化方法的讨论

现有的采样策略大部分是以样本信息量作为挑选标准[10,23,43,5]，这种方法

的主要缺点在于，均没有利用到大量的未标注数据集，并且样本的挑选可能会局限于某一局部范围从而造成抽样偏差。鉴于这些缺陷，研究人员提出了一种新的方法，从样本代表性的角度出发，在未标注样本集中选择那些具有代表性的样本，此类方法认为具有代表性的样本周围会聚集大量的样本。基于这种思路，[46,32,12,24]等文献都通过聚类的方法来挖掘代表性强的样本。然而，这些方法虽然考虑到了数据自身潜在分布规律来挑选具有代表性的样本，但这类方法的缺点在于模型的最终性能很大程度取决于聚类的准确性。

有些主动学习算法试着将两种标准（信息量和代表性）进行结合，如文献[46]中，作者提出一种利用聚类信息和分类器边缘信息相结合的挑选策略，该方法的局限在于仅利用分类边缘附近的样本聚类信息，而忽略了未标注样本整体的聚类结构信息。在文献[12]中，作者动态地寻找样本不确定性和多样性的平衡从而挑选待标注样本进行标注。这些方法均特别指定信息量和代表性的结合方式来进行采样，导致模型的次优性能。

根据上述方法中的缺陷，文献[18]使用基于极小极大理论的主动学习挑选策略，同时考虑样本密度和边界分布的信息，中和信息量大和代表性强的样本两种挑选原则，其优点是即使存在噪声的情况下，依旧能保持较高的样例挑选准确性，并且基于大量未标注数据集对样本的代表性进行估计，避免聚类算法对模型性能的限制。文献中的实验证明[18,54]，所提出的 QUIRE 算法是现阶段基于不确定度的采样策略中性能最优的算法。本文将对 QUIRE 的采样策略进行改进，目的是将众包更好地融入主动学习的分类模型中。

2.3 基于众包的主动学习模型

通过之前的介绍，我们已经了解到众包适用于解决机器难解决的问题，例如文本情感分析[33,21]，图像标记[44]、相同实体匹配[3,6,28,34,45]等。当面对有预算限制的实际问题时，我们引入了主动学习，利用基于众包的主动学习模型来提高模型准确率的同时尽可能降低数据标注成本。然而，目前的基于众包的主动学习模型的优化工作存在局限性，这些工作[47,50]通常局限于一个阶段或一个步骤，

他们对众包阶段的标签质量和主动学习采样阶段的样本质量分别进行优化，并没有将两种技术很好地结合。这种现象其实源于这样一个问题，即如何更好地将人群融入模型中？文献[13]中也提到，数据众包的未来工作中有一条是研究如何更好地将众包与主动学习相结合。

与传统的主动学习模型不同，基于众包的主动学习模型将标注任务交由在线未知大众而非领域专家进行标注。由于未知大众的标注能力不一，为了尽可能减少错误标签我们需要对众包进行质量控制。现有的众包质量控制方法中，有的通过衡量人群的标注能力从而挑选答题准确率高的人群来完成任务[31,11]，有的从众包反馈结果中挑选高质量的标签[29,30,11,19]，这些方法均将所有任务都指派出去。然而，结合主动学习模型后的众包任务与传统的众包任务存在本质上的不同，前者可以从所有未标注数据集中挑选一部分样本进行标注，后者则将所有未标注数据集不经筛选统统交给众包进行标注。另外，现有的基于众包的主动学习模型的优化研究并不多，文献[47,50]中仅是用主动学习来挑选合适的众包工作者，他们将众包阶段和样例选择阶段分离开来，分别进行优化，并没有将众包收集的信息很好地融合到主动学习样本挑选策略中。因此，本文对样本挑选过程进行一定程度的优化，通过融入众包过程中获取的知识，挑选更合适人群的样本来进行众包标注，以此来提升众包标签的质量、改善模型性能。从而将众包更好地融入主动学习模型中。

2.4 本章小结

本章节介绍了现有相关工作及其缺陷，其中首先介绍了众包技术的优势，分析了现有众包反馈形式的优缺点；其次介绍了主动学习的概念，并分析了现有的基于不确定性的样例选择算法的缺陷及其优化方法；最后介绍了基于众包的主动学习模型的相关研究工作，并分析了这些工作所存在的缺陷。接下来，进入本文第一个核心环节——基于众包解释性反馈的主动学习模型优化方法。

第三章 基于众包解释性反馈的主动学习模型优化方法

3.1 方法的引出

本文将在情感极性判断任务中探索我们的方法。例如，在情感分析领域中，判断文本情感极性对于机器而言是困难的，而对于拥有基本阅读能力的人而言，则是比较容易的。既然如此，我们可以从人出发，试图从人身上获取更多的信息，交授给机器。下面展示一些真实微博评论，如表 3.1 所示。

表 3.1 微博评论数据分析

	微博评论	关键词	分析该语境下的含义
网络用语	杯具啊!!! 我的手机又不能开机了，肿么办，修都修不好了，看来你的命数已尽，去死吧!	杯具、肿么办	悲剧（谐音）、怎么办（谐音）
	碉堡了。。。哈哈哈 三星好啊。。。以后可以上厕所带手机了哈哈哈。	碉堡	表示极度震惊或非常厉害的意思。
	我说实话 翡翠真是不好看 表打我 我说的是实话。	表	不要（连音）
特殊含义	这个电影真垃圾!!	垃圾	很烂、差劲 （垃圾原本只表示无用的东西，没有情感色彩）
	马克！待会儿回来看！	马克	英文 Mark 的中文音译，标记的意思。 （表达对转发内容很感兴趣，原本没有这层含义）
正词反用	她就是个白莲花。[微笑]	白莲花	看上去如天使般纯洁,其实内心阴险的人 （带有讽刺意味）
	中国平安不卖保险了，买的是综合金融理财服务~~~搞笑！	搞笑	可笑（带有嘲讽语气）
不同词语在同一句话中的重要程度不同	整体还行，可是价格不能忍！	还行、不能忍	句中有两种情感色彩的词汇，而整体情感色彩是负面的，可见“不能忍”一词在句中的影响度更大。

我们通过分析微博评论中的关键词，发现微博数据的四个特点：1）评论中存在大量的网络新兴词汇。2）在特殊语境下，一些原本没有情感色彩的词语被赋予了新的情感。3）人们经常会用正词反用方式来讽刺某个人或事。4）不同词语在用一句话中的重要程度不同。随着互联网时代的高速发展，会不断孕育出新

的词汇，不同的表达方式也会不停地更新变化。机器在训练集不够的情况下无法顺应这种变化，更无法解决文本中的正词反用、特殊含义等问题。然而人则不同，凭借着先天的优势，能够快速直接地判断出词语的正确情感色彩。无论是遇到网络用语、正词反用或是特殊含义等不同的情况，人们均能感知到语言背后的情感。因此，如果我们能够在众包标注任务过程中挖掘到更多人对数据的潜在认识并将其融入机器学习模型中，那么便能以较少的训练数据得到一个较优的模型。

目前众包平台的标注任务通常仅收集数据标签，这种方式没有充分发挥众包平台的交互作用，众包平台反馈的内容过于贫乏单一。虽然有一些新型的众包形式（如设计一款游戏），然而这类方法太过复杂不易于实现。本文针对现有众包标注任务的局限性，提出了一种基于众包解释性反馈的主动学习模型优化方法。这一方法主要面临以下三个挑战：

1. 如何设计众包任务，在尽可能不影响用户体验的情况下，启发工作者给出更多有用的知识。

2. 如何设计反馈形式，使得收集的反馈内容能够更方便地处理并融入模型中，最终提升模型性能。

3. 让用户提供更多信息意味着要提供更多的报酬来吸引用户挑选任务，那么准确率和成本应该如何平衡。

为了克服上述挑战，本文从任务设计的角度，设计一种众包任务以扩充众包收集的信息——让用户提供标签的同时，给出解释性信息，挖掘用户对数据的潜在理解能力，并且将其融入主动学习模型，迭代地优化模型。

接下来，我们先介绍概念定义和方法框架，然后在情感极性判断任务中探索本文的方法。

3.2 概念定义及方法流程框架

3.2.1 众包反馈与解释性反馈

众包反馈 本文将众包反馈定义为任务发布者通过众包平台收集的内容。对

于不同的任务不同的问题形式，收集内容的形式也可以不同，这往往取决于任务设计方式。本文的众包标注任务中，收集的反馈包括数据标签和解释性反馈。数据标签即为数据所属的类别。以短文本情感分析任务为例，工作者需要判断出短文本的情感倾向，如正面或负面。

解释性反馈 基于本文作者已经阅读过的文献，“解释性反馈”是本文首次提出的概念，本文将其定义为工作者对给出的标签提供的解释，通俗地理解就是回答理由。对于许多机器难以处理的问题，例如情感分析、图像识别等，工作者往往能直观地给出分类标签，然而，这类任务的回答理由往往是工作者潜意识产生的，试图挖掘这种潜意识是艰巨的，可见如何设计任务引导工作者给予所期望的信息是一个巨大挑战。由于数据本身包含了大量的语义信息，我们希望在标准数据集不充足的情况下，尽可能学习到数据本身对分类结果影响最大的一部分，因此我们将数据进行切割，让用户挑选最能影响其作出分类判断的局部内容作为解释性反馈进行收集。解释性反馈包含用户对数据更多的理解性知识，可以预见的是，如果能够将其融入到模型中，那么模型便能更理想地模拟人类行为，对一些机器难处理的问题有更好地预测。

例如，情感极性分析中，若文本中有“开心”一词的出现，该文本往往就是正面的，“开心”的出现对于文本分类有决定性作用，并且它是原文本中的一部分。再例如，图像标注任务中，如图 3.1 所示，当我们要区分狼和哈士奇时，我们能够直观地区分这两种动物通过观察图像中的局部内容，他们的眼睛和尾巴。然而，机器做不到这一点。

图 3.1 图像识别任务识别狼和哈士奇实例

3.2.2 方法流程框架

本文提出基于众包解释性反馈的主动学习模型优化方法，方法流程框架如图3.2 所示。

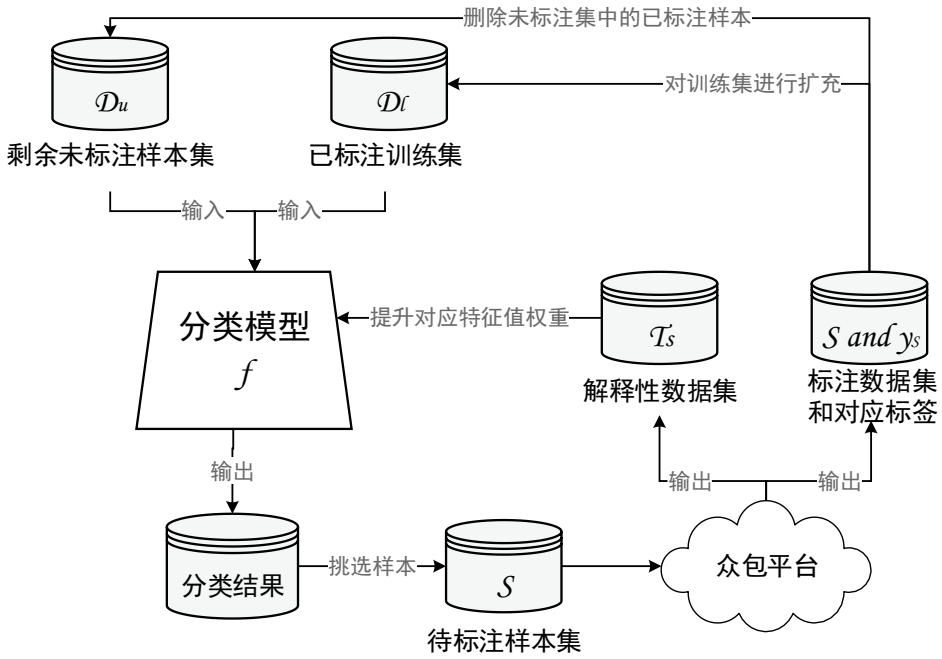


图 3.2 基于众包解释性反馈的主动学习模型框架

模型的具体步骤如下：

1. 训练初始模型。根据少量初始训练集 D_l 训练初始模型 f ，

2. 挑选标注样本。利用主动学习挑选策略挑选机器最不确定的样本集交由众包平台

3. 收集众包反馈信息。根据本文提出的任务设计方式发布数据标注任务（详情见 5.3 节），收集众包反馈信息 $<S, y_s, T_s>$ ，其中包括标签信息 y_s 和解释性信息 T_s 两部分内容。

4. 更新训练集与模型输入。将标签信息 $<S, y_s>$ 加入到训练集 D_l ，同时从剩余未标注数据集 D_u 中删除已标注数据集 S ，并利用解释性信息 T_s 更新模型输入，重新训练模型 f 。

5. 重复 2，3，4 步直到准确率满足预设要求或众包预算用完。

接下来，我们将详细描述众包任务设计方法和融入解释性反馈的模型优化

方法，在第五章通过实验验证我们的猜测。

3.3 众包解释性反馈的收集及主动学习模型改进方案

3.3.1 众包解释性反馈的收集

任务发布者可以通过设计不同的任务，收集不同形式的信息。在初始训练集不充足的情况下，让工作者给出更多对数据的解释性信息，对提升模型性能大有帮助。从众包平台收集反馈的方式有很多种，最直接的方式便是让他们直接写出答案理由，例如：

“你选择该选项的理由是什么？”

让他们写出纯文本的回答理由。然而，在实践过程中，我们发现这样收集反馈的缺陷有以下几点：1）问题太宽泛，解释性的文本没有任何限定，工作者可能给出各种不同的内容，甚至答非所问骗取奖金。2）用户体验不佳。工作者需要消耗更多脑力和精力思考并编辑文本，这会增加他们的工作量，消耗更多预算，不仅如此，答案质量也无法保证。例如，在一些问卷调查中，对于一些填空题，人们往往倾向于填写尽可能少的文字，或者只输入一些标点符号。因为人们总是倾向于被动接受信息，偏爱完成一些有选项供选择的问题。3）工作者提供的文本信息难以处理。不同的人会用不同的词汇去描述相同的事物，这会增加数据处理的难度。

为了解决这些困难，本文设计了一种能够收集规则反馈的众包任务。我们将解释性反馈从形式上分为两类，一类是不规则反馈，此类反馈指的是未限定内容的自然语言；另一类是规则反馈，是事先限定内容范围的信息，可以是原始数据本身的一部分。人们提供的标签也是一种规则反馈。由于人们更倾向于做选择题类型的问题，我们设计了一个交互形式，事先设定好规则的反馈信息，这里我们规定反馈信息为原始数据的局部，然后让工作者给出最能影响其判断的某个或若干个反馈。例如在短文本情感极性判断任务中，解释性反馈就是文本中出现的词语；图像标注任务中，解释性反馈就是图像的局部区域。以微博情感极性判断任

务为例，先让工作者对数据进行标注，然后让他们回答这样的问题：

“以下出现在文本中的词或短语，哪一项或多项选择会让你产生之前的情感判断？”

如图 3.3 所示是众包任务界面的一个示例。这个任务是由两道题组合而成，工作者只有完成第一题后，才能进入第二题。进入第二题后，工作者可以勾选多个词作为该任务的解释性反馈。

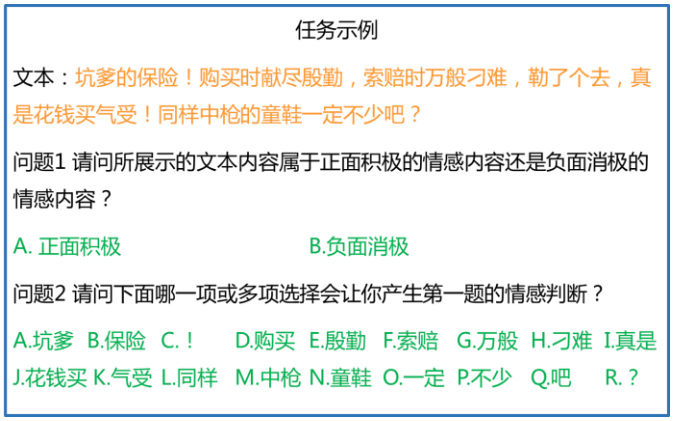


图 3.3 众包任务界面示例

最终，我们收集两部分内容，一是数据标签，二是规则的解释性反馈。这种方法比先前提到的直接提问的方法更好，因为 1）它挖掘到工作者潜意识里的信息，获取重要的数据特征；2）用户体验良好，不用用户自己空想，避免文本框形式的填充，减少工作者脑力消耗以及繁琐的操作；3）规则的解释性信息便于处理。每个解释性信息相当于模型输入空间的一维特征，可以直接通过改变特征对应的权重来将工作者提供的潜意识知识融入模型中。

3.3.2 融入众包解释性反馈的主动学习模型

为了从人身上获取更丰富的知识，本文提出了一种基于众包解释性反馈的主动学习模型，用上一节提出的任务设计方法收集用户解释性信息，挖掘用户对数据潜在的理解性知识，并且将其融入主动学习模型，迭代地优化模型。本节将给出问题的形式化定义以及算法的描述。

假设给定的数据中共有 n 个样本 $D=\{(x_1,y_1),(x_2,y_2),...,(x_{n_l},y_{n_l}),x_{n_l+1},...,x_n\}$，包含 n_l 个黄金标准样本和 $n_u=n-n_l$ 个未标注样本，其中每个样本

$x_i =< x_{i1}, x_{i2}, ..., x_{id} >^{\mathrm{T}}$ 由 d 维特征向量表示，对应的标签 $y_i \in \{-1, +1\}$，处理一个二分类问题。$S=\{x_{s_1}, x_{s_2}, ..., x_{s_k}\}$ 表示主动学习模型中挑选出的待标注样本集，其中 $k \geq 1$ 表示批量挑选样本个数。整个数据集 D 由黄金标准数据集 D_l 、待标注数据集 S 和剩余未标注数据集 D_u 三部分组成，$D = D_l \cup S \cup D_u$ 。我们用 $D_a = D_u \cup S$ 表示所有没有标准答案的样本，用 $y = [y_l, y_s, y_u]$ 表示所有数据标签，并与 D_l 、S 、D_u 中的样本一一对应。同样地，定义 $y_a = [y_s, y_u]$ 为所有未知的数据标签，与 D_a 对应。

基础主动学习算法由以下五个组件构成：分类器 f；初始标准训练集 D_l；采样策略 Q；督导者，即众包 $Crowd$；未标注数据集 D_a 。我们的目标是优化模型性能得到尽可能多的正确标签，我们的具体做法是，将主动学习算法 Q 挑选出的样本集 S 交由众包平台 $Crowd$，收集平台返回的标签反馈 y_s 和解释性反馈 T_s，其中 $T_s = \{t_i \mid t_i \in T\}$，$T = \{t_1, t_2, ..., t_d\}$ 是 X 的特征集合。然后，将 $< S, y_s >$ 扩充入训练集中，将 T_s 以增加对应权重的方式融入到分类器 f 中，其中记 $W =< w_1, w_2, ..., w_d >$，$w_i$ 对应每个特征的权重系数，预测剩余数据集，这里需要定义两个参数变量 C 表示影响因子、I 表示权重更新次数。反复迭代，直至预算 B 用完或准确率达到阈值 A。最终，我们返回所有没有标准答案的标签 $y_a = [y_s, y_u]$，即包含众包标签和模型预测标签两部分。

用 $< S, y_S, T_S >= Crowd(S)$ 表示众包平台返回的标签和解释性关键词集合。为了避免一些不确定因素影响收集标签的质量，例如用户点错按钮，我们运用了基于冗余信息的质量控制策略。每道题我们收集三个答案，并选用多项投票原则将多数人选的标签赋值给 y_S，更新训练集 $D_l = D_l \cup (S, y_s)$，并将选择 y_S 标签的用户对应的关键词集合记录在 T_S。

接下来，通过改变模型输入将解释性反馈融入模型中。$T = \{t_1, t_2, ..., t_d\}$ 表示微博数据集的特征集合，样本集 $X = \{x_1, x_2, ..., x_n\}$ 每个样本 $x_i =< x_{i1}, x_{i2}, ..., x_{id} >$ 表示 T 中对应特征的特征值。定义 $W = [w_1, w_2 ..., w_d]$ 为 T 中对应特征的权重值，所有权重初始值均为 1。

$$W = \begin{cases} w_i = C * w_i & if\ t_i \in T_S \\ w_i & else \end{cases} \tag{3-1}$$

利用公式 3-1 对 W 进行更行，若特征t_i存在在解释性反馈集合T_S中，则对对应的权重乘上一个权重系数 C，增加它的权重值，利用权重向量 W 对输入词向量 X 进行更新，更新公式如 3-2 式所示：

$$X = X^{\mathrm{T}} \times W \tag{3-2}$$

将更新后的词向量 X 输入模型，重新训练分类器。反复迭代，直到准确率满足预设要求或众包预算用完。

算法 3.1，展示了基于众包解释性反馈的主动学习模型优化方法。

算法 3.1 基于众包解释性反馈的主动学习模型算法

Input：D，B，A // 完整数据集，预算，准确率阈值

Output：y_a // 所有没有标准答案的数据标签

Initialize:

$D_l = \{(x_1, y_1), (x_2, y_2), ..., (x_{n_l}, y_{n_l})\}$ // 初始训练集

$D_u = \{x_{n_l+1}, ..., x_n\}$，$S = \varnothing$ // 未标注数据集，待标注样本集

$X = \{x_1, x_2, ..., x_n\}, x_i =< x_{i1}, x_{i2}, ..., x_{id} >$ // 模型输入

$W =< w_1, w_2, ..., w_d >$，$w_i = 1$ // 权重系数

l, n, k // 已标记样本数、总样本数、每次挑选样本数

Parameter：C，I // 权重系数、权重更新轮次

1:**Repeat**:

2: $f^* = Model\_Train(D_l)$ // 训练分类器

3: $S = \underset{S \subseteq D_u \wedge |S| = k}{\arg\max} Q(D_u)$ // 用主动学习算法挑选 k 个信息量最大样本

4: $< S, y_S, T_S >= Crowd(S)$

/*获取众包反馈信息，包括标签信息和解释性信息*/

5: $D_l = D_l \cup (S, y_S)$，$D_u = D_u / S$ // 更新训练集和未标注样本集

6: **While** ($I > 0$):

7: For $t_1 \sim t_d$ in T_S:

/*更新存在于解释性反馈中的特征的对应权重系数*/

8: Put $w_i = C \cdot w_i$

9: $X = X^{\mathrm{T}} \cdot dia\_matrix(W)$

10: $I--$

11:**Until** the budget B and accuracy A is reached.

12: $y_u = f^*(D_u)$ // 用最终的分类器对剩余未标注数据进行预测

13:**Return** $y_a = [y_S, y_u]$

14:**End**

3.4 本章小结

本章首先通过分析一组真实环境下的数据集的特点，引出本文提出的方法的主要思想，然后介绍了方法中涉及的相关概念和具体步骤，接着详细描述了众包解释性反馈的收集过程，并在情感极性判断数据集上进行了方法的探索与运用。

第四章 融入众包标注置信度的主动学习算法改进方法

4.1 方法的引出

文献[30]中提出，同一任务下的不同样本标注难度是不一致的。他们利用CMU 人脸图像集进行众包实验，其目的是判别图像中人的情绪（开心、悲伤、生气），实验中将人脸图像根据不同朝向进行分组，实验结果如表 4.1 所示：

表 4. 1 人脸情绪识别实验结果

Facial orientation	**Avg.accuracy**
straight	0.6335
left	0.6216
right	0.6049
up	0.4805

表 4.1 中第一列表示人脸的不同朝向（向前、向左、向右、向上），第二列表示不同分组下对应的平均众包标注准确率。由表中数据可见，众包人群对同一任务下的不同样本的标注能力不一。如果将不合适的样本交由众包标注收回的标签错误率大，势必会影响模型性能。因此，提出我们的合理设想：在挑选样本过程中，挑选合适大众标注的样本能够减少众包标签的错误率，优化模型输入从而提升模型分类性能。

然而仅从人的角度来衡量样本的优劣是不合理的，这只是个局部视角，还应该将样本信息量、代表性放在一起综合地考虑。由此引出我们的挑选策略的主要思想：在保证样本充足的信息量和代表性的同时，挑选适合大众标注的样本进行标注，从而减少众包标签错误率。本方法主要面临的挑战如下：

1. 如何衡量一个样本是否适合交由众包平台进行标注。
2. 如何将新的挑选原则，融入原有的主动学习采样原则中，从而实现本文提出的挑选思想。

为了克服这些挑战，我们在文献[18]中提出的采样策略 QUIRE 算法的基础上进行改进，融入一种新的挑选标准——众包标注置信度，综合考量信息量、代表性和标注难度三种标准来挑选样本，使得众包和主动学习两种技术能够更好地

结合。

4.2 概念定义及方法流程框架

4.2.1 众包偏差与众包标注置信度

众包偏差 在运用众包技术来执行数据标注任务时，由于许多未知因素影响，例如工作者能力、任务难度或者任务描述等，会导致收集到的标签存在一些噪音和错误，这一现象，我们定义为众包偏差，这反映的是众包反馈的标签质量问题。直观地理解，如果众包给出的标签是有偏差的，那么模型势必会被错误标签误导，导致性能的下降。

我们对现有方法进行改进前，需要引入一个新的衡量标准——**众包标注置信度**。我们要从样本中挑选众包标注置信度高的样本，即适合人群的样本，进行标注，从而提升标签质量。每个样本对应一个众包标注置信度。我们猜测，在主动学习挑选样本过程中，在考虑样本本身信息量和代表性的同时，加入众包标注置信度的考量后，能够一定程度降低众包偏差，提升众包标签的质量从而优化输入，达到提升模型性能的最终目的。接下来我们将详细介绍方法的流程，在第五章通过实验来验证我们的猜测。

4.2.2 方法流程框架

本文基于文献[18]中提出的主动学习挑选策略进行优化，融入众包标注置信度，综合考量样本信息量、样本代表性和样本标注难度三种衡量标准，提出一种能够更好地结合众包与主动学习的模型框架，算法优化流程如图 4.1 所示。

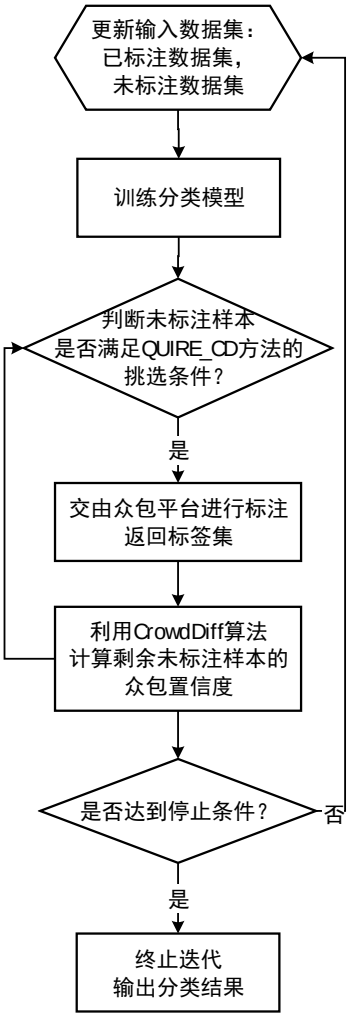


图 4.1 融入众包标注置信度的主动学习挑选策略流程图

本文融入众包标注置信度的主动学习算法优化方法 QUIRE_CD 的整体流程描述如下：

由图 4.1 所示，整个优化算法的输入包括两部分：已标注数据集、剩余未标注数据集。整个 QUIRE_CD 优化方法的核心在于挑选样本时加入样本众包置信度的衡量标准：基于文献[18]中提出的 QUIRE 策略基础上进行改进，提出一个 CrowdDiff 算法估计众包标注置信度，综合考虑样本信息量和代表性的同时，挑选更适合人群的样本集，交由众包平台进行标注。这个过程反复迭代，直到达到算法停止条件。我们猜想，改进后的样例选择算法能够综合三种衡量标准，并挑选出最优的待标注样本集。

接下来，我们将详细描述融入众包标注置信度的主动学习算法具体优化方法，

并在第五章通过实验来验证我们的猜想。

4.3 融入众包标注置信度的采样策略改进方法

4.3.1QUIRE 算法介绍

这一小节回顾基础方法 QUIRE 样本挑选策略[18,55]，该方法与传统方法不同之处在于，它综合利用少量已标注数据集和大量未标注数据集，更好地结合了样本信息量和代表性两方面的特性进行挑选，从而提高分类器的性能。

文献通过理论化的分析与证明，提出一个样本挑选的评估函数如 4-1 式所示：

$$\hat{L}(D_l, D_u, x_s) = \min_{y_u \in \{\pm 1\}^{n_u - 1}} \max_{y_s = \pm 1} y^{\mathrm{T}} ([\kappa(x_i, x_j)]_{n \times n} + \lambda I)^{-1} y \qquad (4\text{-}1)$$

其中 $D_l = \{(x_1, y_1), (x_2, y_2), ..., (x_{n_l}, y_{n_l})\}$， $D_u = \{x_{n_l+1}, ..., x_n\}$ 且 $x_s \in D_u$ 。

文献通过数学推导，将评估函数近似地写成两项之和，其中第一项用已标注样本集预测x_s的置信度，对应于它的信息量；第二项利用大量未标注样本集预测x_s与未标注数据的契合程度，对应于它的代表性。该挑选策略能够在多种数据集上保证挑选出的样本质量，文献[18]通过实验论证了该方法是现有的基于不确定度采样策略中性能最好的样例选择算法。

4.3.2 融入众包标注置信度的采样策略——QUIRE_CrowdDiff

首先，我们详细介绍众包标注置信度的计算方法 **CrowdDiff**。基于文献[53]中提出的理论依据，我们首先利用已标注样本标签的离散度来估计众包对已标注样本的标注能力，然后利用已标注样本估计值估计每个未标注样本的众包标注置信度。如果一个未标注样本与已标注样本相似性越大，我们认为众包对这两个样本的标注能力越相近。形象地说，离样本越近的已标注样本对该样本的影响越大。我们用欧氏距离来衡量样本之间的相似度，并通过加权的方式估计未标注样本的众包标注置信度，即 **CrowdDiff 值**。计算流程如图 4.2 所示。

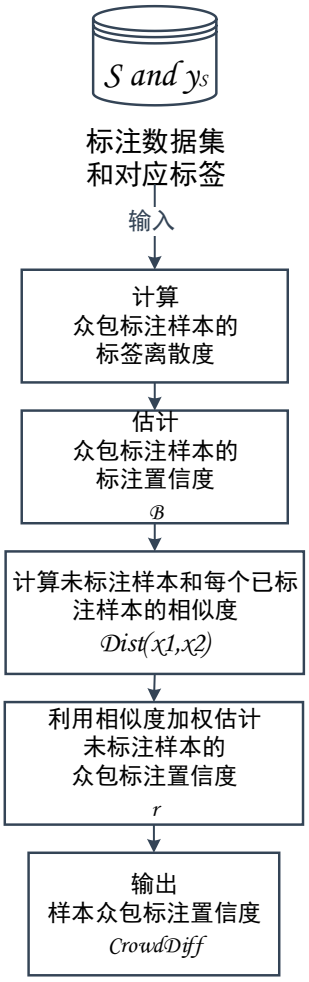


图 4.2 样本众包标注置信度算法流程图

首先，我们通过估计已标注样本的离散度来估计其众包标注难度。这里我们参考文献[53]中提出的样本平衡度的计算方法来计算标注样本离散度。

沿用本文第三章中的形式化定义，再定义：

$L=\{L_i^j \mid i=1,2,...,|S_{all}|; j=1,2,...,m\}$ 其中 L_i^j 表示 S_{all} 中第 i 个样本 x_i 的第 j 个标签，每个样本收集 m 个标签，S_{all} 记录所有众包标注样本，每次回收到众包标签反馈时自动更新 L。根据公式 4-2 计算每个选项的支持票数，即多少人选了这个选项。

$$N_i^q=\sum_{j=1}^{m} I(L_i^j=Query_q) \quad \text{，其中I(x)} = \begin{cases} 1 & \text{若 x 为真} \\ 0 & \text{若 x 为假} \end{cases} \quad \text{，} \quad q=1,2,...,Q \quad (4\text{-}2)$$

N_i^q 表示样本 x_{s_i} 的众包任务中第 q 个选项的支持票数，总共有 Q 个不同选项。然后用支持票数/总票数得到每个选项的支持率，如下式 4-3 所示。

$$R_i^q = \frac{N_i^q}{m} \tag{4-3}$$

用方差公式计算每个选项的支持率离散程度，由此来反映已标注样本的标注难度，如下式 4-4 所示。

$$\beta_{x_i} = \frac{\sum_{q=1}^{Q}(R_i^q - \bar{R}_i)^2}{Q} + 1 \tag{4-4}$$

为了利用已有的标签信息估计众包对未标注样本的标注能力，根据我们先前提出的假设，我们基于已标注样本的标注难度，利用样本之间的相似度来加权估算未标注样本的标注难度，这里的相似度用欧氏距离来表示，如下式 4-5 所示。

$$\gamma_{x_j} = \sum_{i=1}^{|S_{all}|} \frac{\beta_{x_i}}{dist(x_i, x_j)} \tag{4-5}$$

其中 $dist(x_i, x_j) = \sqrt{\sum_{v=1}^{d}(x_{iv} - x_{jv})^2}$ 用欧氏距离来表示。

$dist(x_i, x_j)$ 表示众包已标记样本 $x_i \in S_{all}$ 与未标注样本 $x_j \in D_u$ 的距离，γ_{x_j} 表示未标注样本利用已收集的众包标签估计而得的标注难度。

得到未标注样本标注难度即可反映众包对样本的标注置信度，标注难度 γ_{x_i} 越大众包置信度越低，用公式 4-6 来表示每次挑选出的样本整体众包标注置信度。

$$CrowdDiff(L, D_u, S_{all}, S) = \sum_{x_j \in S} \gamma_{x_j} \quad （x_j \in D_u） \tag{4-6}$$

结合公式 4-1 得到最终的挑选策略可以表示成 4-7 式：

$$S = \underset{S \subseteq D_u \wedge |S| = k}{\arg\min} [(1-t) \cdot \hat{L}(D_l, D_u, S) + t \cdot CrowdDiff(L, D_u, S_{all}, S)] \tag{4-7}$$

其中 t 表示样本标注难度的影响因子。t 越大表示在挑选过程中对公式 4-7 中对第二项的约束更严格，反之则约束更小，其取值将在实验部分讨论。

融入样本标注难度的主动学习优化方法 *QUIRE _CrowdDiff* 算法步骤如下算法 4.1 所示。

算法 4.1 融入样本标注难度的主动学习算法改进方法

Input： D_l， D_u // 完整数据集

$L=\{L_i^j \mid i=1,2,...,|S_{all}|; j=1,2,...,m\}$

// 众包平台收集的标签集合，初始值为空。

// $S_{all}=\{x_i|$所有交给众包标注的样本集$\}$， m 表示每个任务收集答案数

Output： y_a // 所有没有标准答案的数据标签

Initialize:

$D_l=\{(x_1,y_1),(x_2,y_2),...,(x_{n_l},y_{n_l})\}$ // 初始训练集

$D_u=\{x_{n_l+1},....,x_n\}$ ， $S=\varnothing$ // 未标注数据集，待标注样本集

l,n,k // 已标记样本数、总样本数、每次挑选样本数

Parameter： t // 影响因子，默认为 0.5

1:**Repeat**:

2: $f^*=Model\_Train(D_l)$ // 用有标签训练集训练分类器

3: $S=\underset{S\subseteq D_u \wedge |S|=k}{\arg\min}[(1-t)\cdot \hat{L}(D_l,D_u,S)+t\cdot CrowdDiff(L,D_u,S_{all})]$

// 用公式 4-7 选 k 个样本交给众包平台

4: $<S,y_S>=Crowd(S)$ // 获取众包反馈标签信息

5: $D_l=D_l\cup(S,y_s)$， $D_u=D_u/S$ // 更新训练集和未标注样本集

6: $S_{all}=S_{all}+S$ // 更新众包已标注样本集

7:**Until** the budget B and accuracy A is reached.

8: $y_u=f^*(D_u)$ // 用最终的分类器对剩余未标注数据进行预测

9:**Return** $y_a=[y_s,y_u]$

10:**End**

4.4 本章小结

本章节是在现有的主动学习挑选策略 QUIRE 基础上进行改进，融入众包标注置信度的考量，目的是在保证样本信息量和代表性的同时，尽可能地挑选适合众包标注的样本，较少众包噪音给模型带来的负面影响。本章节通过对真实环境下的实验数据进行分析，引出本文方法的主要思想，然后详细介绍相关概念和方法框架，并提出了众标注置信度的计算方法，最后展示整个优化方法的执行算法。

第五章 实验设计与结果分析

5.1 众包过程

我们选择阿里众包作为收集数据信息的众包平台，阿里众包对工作者条件有一定的限制（包括实名认证、信用分大于 550 分等），因此本文假设该平台能保证一定的标签质量。众包平台发布标注任务前，需要事先设计众包任务。如图 5.1 所示，展示了本文实验中微博情感极性标注的任务描述、任务示例以及单价信息。

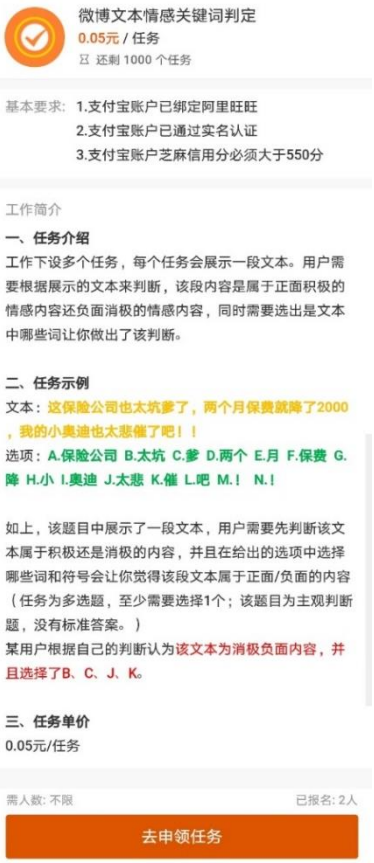


图 5.1 微博情感极性判断任务信息

由图 5.1 可见，任务描述中限制了工作者的身份，介绍了任务是由多个问题组成，并且展示了一个任务示例，让工作者更清晰地了解到任务目的。经过市场调查后，我们将任务单价定为 0.05 元/任务。我们每道题收集 3 个答案，用多数投票原则进行一定的质量控制，即选择大多数人所挑选的标签作为最终答案，加入训练集中。

根据 3.3.1 节的反馈收集方式介绍，可知每个任务包含两道题，一题是获取标签，如图 5.2(a)；另一题是获取解释性信息，如图 5.2(b)。

(a) 微博情感极性判断　　(b) 挑选解释性反馈

图 5.2 众包任务示例图

下面列举了部分众包反馈的标签和解释性数据，如表 5.1 所示：

表 5.1 阿里众包反馈示例

Task_id	User_id	情感极性	选项 1	选项 2	选项 3	选项 4	选项 5	选项 6
950	user1	正面积极	哈哈哈	好	啊			
	user2	正面积极	碉堡	哈哈哈	好			
	user3	正面积极	哈哈哈					
709	user1	正面积极	神机					
	user2	正面积极	好	神机				
	user3	正面积极	神机					
2709	user1	负面消极	电话	加	黑名单			
	user2	负面消极	推销	不感兴趣	黑名单			
	user3	负面消极	电话	加	黑名单	不买	任何	保险
4678	user1	负面消极	浮云	妹	坑爹			
	user2	负面消极	搞	浮云	妹			
	user3	负面消极	浮云	妹	坑爹			
4684	user1	负面消极	不敢当	孬货				
	user2	负面消极	孬货	草				
	user3	负面消极	孬货					

表 5.1 中展示了任务 id，情感极性标签，以及解释性信息，同一个 task_id 对应一条微博文本，每条文本收集三个用户给出的答案。我们发现用户挑选出的解释性关键词数量并不多，平均每道题给出 2-3 个解释性反馈。给出的关键词中包

含一些网络用语，如“碉堡”、“神机”、“浮云”等，以及一些特殊含义名词，如“草”不是一种植物而是表达强烈的不满，这些关键词都是用户认为能够影响他们做出情感极性判断的关键词。对于机器而言，在标准训练集不充分的情况下，很难发现这些关键词的重要性，由此可见，我们设计的众包任务能够挖掘到用户对数据的潜在认识。

5.2 评价指标

1．剩余数据集准确率 Accuracy

这里的准确率不单单是模型的预测准确率，还包括一部分众包收集到的标签准确率。计算公式如 5-1 式所示：

$$Accuracy = \frac{n_{tc} + n_{tm}}{n_a} \tag{5-1}$$

n_{tc} 表示从众包中收集到的正确标签数，n_{tm} 表示模型预测正确的标签数，n_a 表示除带标签的初始数据集以外的样本数。

2．精确率 P、召回率 R 和调和均值 F1-measure

先介绍 TP,FN,FP,TN 四种分类情况，如表 5.2 所示：

表 5. 2 四种分类情况

		真实类别	
		正类	负类
预测类别	正类	True Positives（TP）	False Positives（FP）
	父类	False Negatives（FN）	True Negatives（TN）

TP：表示将正类预测为正类的样本数。

FP：表示将负类预测为正类的样本数。

FN：表示将正类预测为负类的样本数。

TN：表示将负类预测为负类的样本数。

精确率反映的是所有预测为正类的样本中真正为正类的比重，计算公式如 5-2 式所示：

$$p = \frac{TP}{TP + FP} \tag{5-2}$$

召回率反映的是被正确预测为正类的样本占所有正类样本的比重，计算公式如 5-3 式所示：

$$R = \frac{TP}{TP + FN} \tag{5-3}$$

当精确率和召回率出现矛盾时，我们可以用 F1-measure 综合考虑问题。F1-measure 就是精确率和召回率的调和均值，如公式 5-4 所示：

$$F_1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN} \tag{5-4}$$

3．标注问题数 N_q

N_q 表示交由众包标注的样本个数，这个指标同样能用来衡量众包任务成本。

5.3 实验设计

1 实验数据集

本文针对情感分析领域中的情感极性判断任务，选取两组数据集进行实验。分别是：酒店评论语料和微博评论文本。两个数据集的相关信息如表 5.3 所示。

表 5.3 实验数据集详情

	酒店评论语料	微博评论文本
总数（条）	10000	4000
正样本数（条）	4000	2000
负样本数（条）	6000	2000
来源	携程网	COAE2014
数据描述	中科院博士谭松波收集整理的酒店评论语料，语料从携程网上自动采集，并经过整理而成	来自第六届中文倾向性分析评测（COAE2014）第 8 个赛题的评测数据集，并提供了正确的情感极性标签。

2 对照实验

1）Baseline：用基于传统众包标签的方法作为实验的基准。这里的主动学习挑选策略为随机采样算法（RND），严格意义上说，RND 不属于主动学习挑选算

法。实验过程中，每次都从未标注集中随机挑选样本交由众包平台，仅收集数据标签，然后将其加入训练集，每次迭代计算相关评测指标。

2）MS(Margin Sampling)：实验中的主动学习算法换做基于不确定度的边缘抽样策略，总是选取距离分离超平面最近的样本作为不确定度最大的样本进行标注，这是最常用的主动学习挑选策略。该方法是基于支持向量机分类器 SVM 进行的。

3）CEF(Crowd with Explanatory Feedback)：本文提出的基于众包解释性反馈的主动学习优化方法，详细过程见 3.3 节，过程中每次迭代计算相关评测指标。

4）QUIRE(Active Learning by Querying Informative and Representative Examples)：文献[18]中提出的一种结合了样本信息量和代表性两方面原则的挑选策略，详细过程见 4.3 节。该方法是基于支持向量机分类器 SVM 进行的。

5）QUIRE-CD(QUIRE with CrowdDiff)：本文提出的融入样本标注难度（CrowdDiff）的挑选策略。该方法同样是基于支持向量机分类器 SVM 进行的。

3 实验设置：

1）实验模式：一步式、迭代式。这里借鉴文献[29]中提出的两种实验场景：

一步式：所有待标注样本的挑选仅基于初始标记数据训练得到的模型，标签统一仅添加一次，将收集到的标签加入训练集中重新训练一次后，对剩余未标注数据进行预测。

迭代式：每次挑选 k 个待标注样本，每次挑选基于上次的训练模型，每次加入标签重新训练模型，直到达到停止条件后输出当前模型对剩余未标注数据的预测结果，k 默认为 10。

2） 初始训练集：设置为总数据集的 0.5%。由于我们的方法是在训练集不充分的环境下进行讨论的，因此我们的初始训练集不能太大，这里我们设置为总数据量的 0.5%。

3） 分类器：这里选择常用的文本分类器包括：支持向量机 SVM、朴素贝叶斯 NB、k-邻近算法 kNN 以及逻辑回归 LR。

4） 主动学习挑选策略：随机采样算法（RND）、基于不确定度的边缘抽样算法（Margin Sampling）。

我们将在接下来的实验中考察本文提出的方法的有效性，并从多个维度出发讨论方法对模型的优化效果。表 5.4 和表 5.5 分别归纳了 5.4 节和 5.5 节中各组实验的相关内容。

表 5.4 基于众包解释性反馈的主动学习模型的优化方法(CEF)实验

实验章节	实验目的	实验模式和数据集	比较内容
5.4.1	验证 CEF 在两种采样策略下的有效性。	环境模式：一步式 数据集：酒店评论数据集	在随机采样(Baseline)和基于不确定度的采样(MS)两种策略下，引入解释性反馈后，分类准确率的前后差异。
5.4.2	验证 CEF 在不同分类模型下的普适性； 验证 CEF 在不同标签分布下的普适性。	环境模式：一步式 数据集：不均匀酒店评论数据集和均匀的酒店评论数据集	不同分类模型下以及不同标签分布下，CEF 方法对模型分类准确率的提升值。
5.4.3	考察 CEF 在不同权重系数下的优化效果。	环境模式：一步式 数据集：酒店评论数据集	不同权重系数下，CEF 模型的分类准确率。
5.4.4	验证 CEF 方法高效的学习能力。	环境模式：迭代式 数据集：微博评论数据集和酒店评论数据集	Baseline、MS 和 CEF 三种方法基于 SVM 分类器下的模型学习曲线。
5.4.5	考察每次迭代挑选不同样本数对 CEF 最终分类效果的影响。	环境模式：迭代式 数据集：微博评论数据集	保持最终标注总数一致的情况下，不同 batch size 对 Baseline、MS、CEF 和被动学习 Passive 的模型分类准确率的影响。
5.4.6	验证 CEF 能够用更少的花费达到相同的准确率。	环境模式：迭代式 数据集：微博评论数据集	Baseline、MS 和 CEF 三种方法在迭代式环境下达到一致准确率所需要的标注成本和时间代价。

表 5.5 融入众包置信度的采样策略改进方法(QUIRE_CD)实验			
实验章节	实验目的	实验模式和数据集	比较内容
5.5.1	验证引入众包标注置信度的采样策略的有效性。	环境模式：迭代式 数据集：微博评论数据集和酒店评论数据集	Baseline、MS、QUIRE 和 QUIRE_CD 方法基于 SVM 分类器下的模型学习曲线。
5.5.2	验证 QUIRE_CD 方法挑选的样本是否适合众包工作者。	环境模式：迭代式 数据集：微博评论数据集	Baseline、MS、QUIRE 和 QUIRE_CD 四种方法中众包收集的标签准确率。
5.5.3	考察众包标注置信度在不同约束强度下对模型分类性能的影响。	环境模式：一步式 数据集：均匀酒店评论数据	在不同影响因子 t 下，QUIRE_CD 方法的分类准确率。

由于随机挑选初始数据集，模型质量受训练数据影响较大，因此每组实验重复 10 次，记录各项指标的平均值，得到模型在当前实验设置下的平均水平。此外，每次更新后的词向量都进行标准化处理，避免单个特征的样本取值相差甚大或明显不遵从高斯正态分布，这里的标准化方法选用—— Z-score 标准差标准化。

5.4 基于解释性反馈的主动学习模型优化方法实验

5.4.1 不同主动学习算法下的讨论

在一步式环境下，对酒店评论数据集进行实验。初始训练集为 10000*0.5%=50 条，然后利用随机采样和基于不确定度的边缘采样策略挑选 20 条样本交由众包标注，收集众包标签反馈和解释性反馈。在把解释性信息融入模型中时，我们对权重系数 C 赋予一个经验值 18，并标准化更新后的词向量。比较两种挑选策略下，融入解释性反馈后剩余数据集准确率，实验结果如图 5.3 所示：

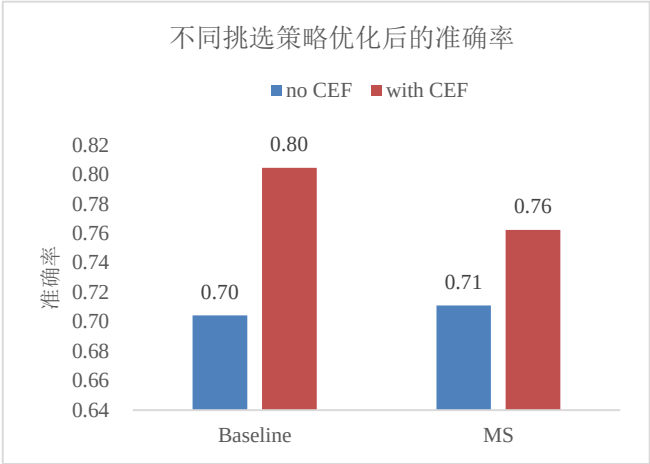


图 5.3 不同挑选策略模型优化后的准确率

由实验数据可见，在随机采样和基于不确定度的边缘采样两种样本选择策略下，加入反馈后模型均有明显提升效果，准确率分别提升 10%和 5%。提升幅度有一定差距，是因为不同的挑选策略导致挑选样本的不同，不同的样本会使用户给出不同的关键词，不同的关键词对模型的影响效果又是完全不同的，一连串变量的不同取值造成了提升效果的不同。虽然准确率提升大小不一致，但依然证明了我们先前的假设：融入解释性反馈后能够给模型性能带来较大的提升效果。

5.4.2 不同分类器和不同标签分布下的讨论

比较不同分类模型加入解释性反馈后的提升效果，以及将该方法运用于不同类别分布数据集后的提升效果。这里，同样地，随机挑选 20 条样本，加入到训练集中。这里的 6000 条数据是由 10000 条数据中正面样本和负面样本分别抽取 3000 条后组合而成，即 6000 条数据的标签是分均分布的，而 10000 条数据集的标签是不均匀分布的。实验中的采样策略均采用基于不确定度的采样策略，且权重系数 C 均默认为 18。实验结果如下图 5.4 所示：

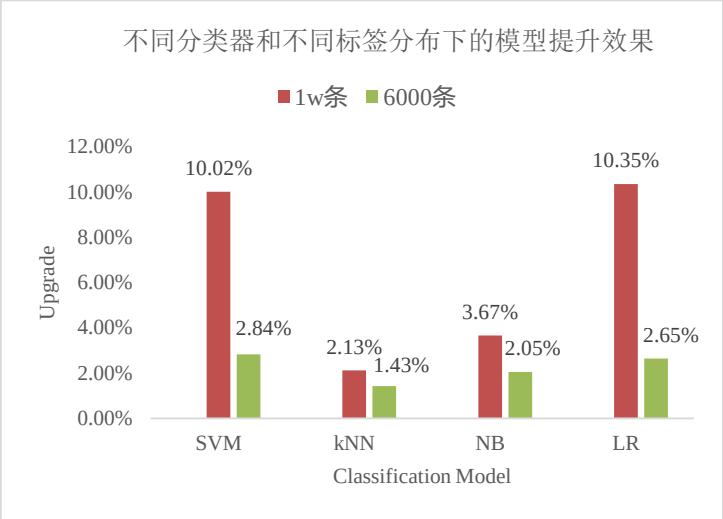


图 5.4 不同分类器和不同标签分布下的模型提升效果

由图 5.4 可见，对于不同分类模型，基于解释性反馈的优化方法对多种模型的分类效果均能起到一定的提升作用，支持向量机 SVM 和逻辑回归 LR 两种模型较其他两种模型的提升效果更好。可见，基于解释性反馈的模型优化方法对于多种分类模型均适用。在之后的实验中，我们在迭代式环境中选择 SVM 算法作为我们的基础分类器。

5.4.3 不同权重系数的讨论

在一步式环境下，在 10000 条酒店评论数据上执行本文提出的 CEF 方法进行实验，以 50 条作为初始训练集随机挑选 20 条交给众包平台收集标签和解释性信息，在融入解释性反馈的过程中设置不同的权重系数，最后观察不同权重系数下的准确率变化曲线。实验结果如图 5.5 所示。

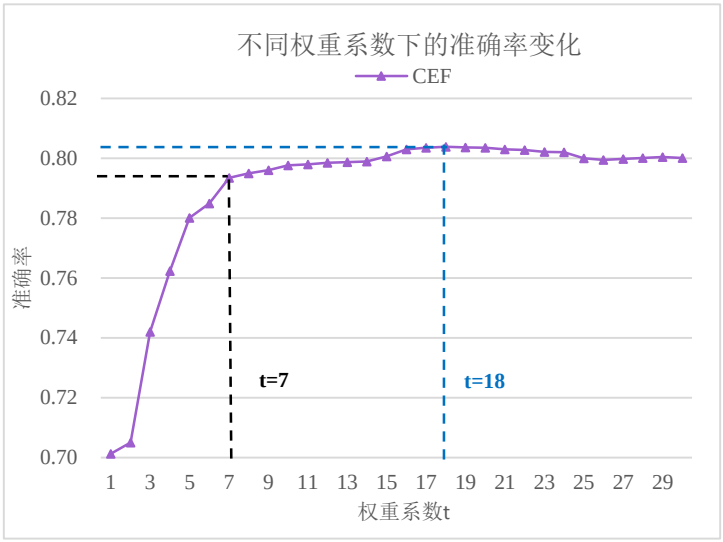
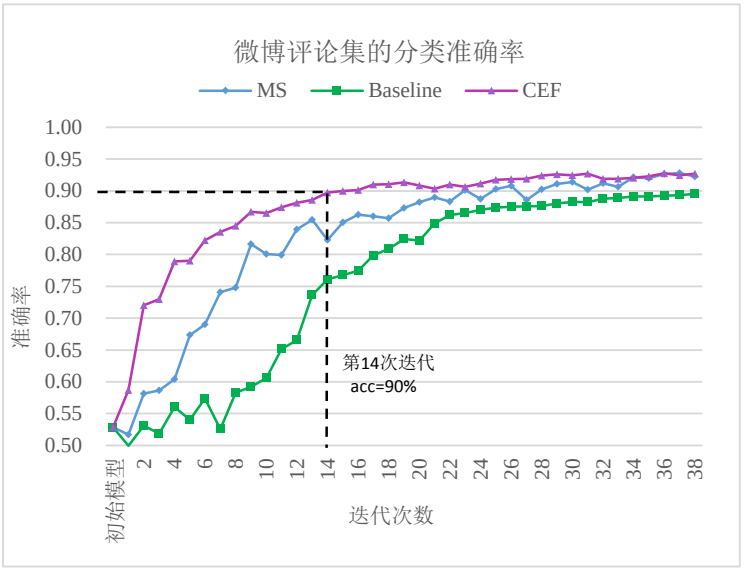


图 5.5 不同权重系数下的准确率变化

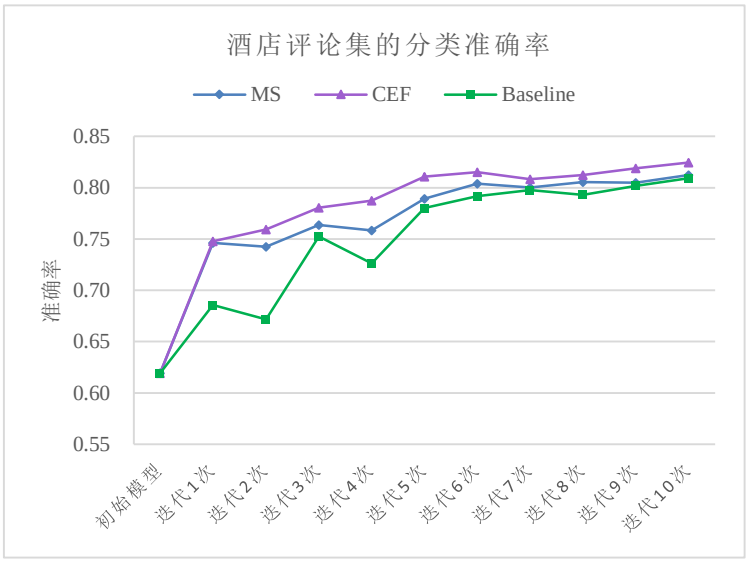
由图 5.5 可以看出，在权重系数 C 为 0-7 的过程中准确率提升效果明显，在 7 以后的提升效果趋于平缓。由此可见，对于用户给出的反馈关键词对模型的提升效果有一定的影响，但当权重系数超过一定范围后，如图所示 t=18 达到最高值，之后会开始下滑。这一现象是符合我们的认识的，每个关键词都有它自己的权重，过高或过低地设置权重值，都会对模型性能造成一定的影响。

5.4.4 迭代式环境下比较不同方法的分类效果

在迭代式环境下，分别比较三组对照实验（Baseline、MS、CEF）在两种评论数据集上的分类效果。三组实验均随机挑选初始训练集：20 条；扩充后的训练集总量不超过数据总量的 10%，即一共标注数量不超过 400 条；每次挑选迭代数量 10 条，共迭代 38 次。迭代式的 CEF 方法中的基础分类模型选定为 SVM，采样策略选定为基于不确定度的边缘采样策略，并且对其中的权重系数 C 和权重更新轮次 I 分别指定为经验值 3 和 10。根据这样的实验设置，观察三种不同方法作用于两种数据集中模型的分类效果，实验结果如图 5.6 所示。



(a) 微博评论



(b) 酒店评论

图 5.6 迭代式环境下不同方法在两种数据集上的分类效果

图 5.6 展示的是不同方法作用在两种数据集上迭代不同次数的模型性能变化曲线，横坐标显示的是迭代次数，0 次表示的是用 0.5%的初始训练集训练的初始模型，每增加一次就是增加 10 条获得标签的数据，微博评论集中一共迭代 38 次，酒店评论集中一共迭代 10 次。由图可见：

i. 整体上看，三种方法的模型学习曲线有明显的区别，本文提出的方法 CEF 和无解释性反馈的方法 MS、Baseline 相比，准确率提升速度更快。图 5.6(a)中 CEF 方法和 MS 方法的最终性能非常接近并且均高出 Baseline 方法 3%左右。更

值得关注的是，CEF 方法是三个方法中学习速率最快的，在前 20 次迭代中，平均提升效率达到 8%，并且从第 14 次迭代开始准确率便达到了 90%，比 Baseline 和 MS 更早地收敛。图 5.6(b)中展现了类似的效果，CEF 曲线在迭代过程中的准确率，比 MS 和 Baseline 方法分别平均提升 1.4%和 3.5%。由此可见，本文提出的方法在一些文本分类问题上，比传统方法效果更好。

ii. 图 5.6(a)中 Baseline 前期准确率非常低 54%左右，且波动比其他方法都大，从第 8 次迭代开始呈现稳步上升状态，原因是由于初始训练集严重缺失，分类模型在数据量有限的情况下，模型分类效果自然很低。另外，随机采样过程中会挑选到一些异常点，影响模型分类效率，因此 Baseline 在迭代初期的模型性能非常差，并且有较大的浮动，效率极其不稳定。随着迭代次数的增加，模型训练集的扩大，模型性能也开始平稳提升，最终趋于平稳。

iii. 由图 5.6 可见，CEF 方法呈现的模型性能是三个方法中最优的，它的曲线较其他两个方法更光滑，并且能更快地到达 90%以上的准确率，最终达到的准确率和 MS 方法几乎持平。由此可得，本文提出的 CEF，基于众包解释性反馈的主动学习优化方法，能够得到更健壮更高效的分类模型。最终，由于模型在标记成本有限的情况下，获取到的信息量也是有限的，所以 CEF 方法最终性能与 MS 方法持平，均近似达到 93%。

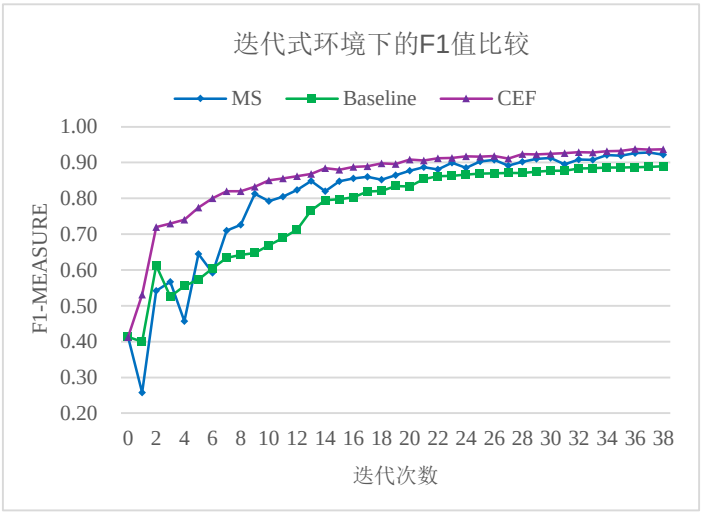


图 5.7 迭代式环境下的 F1 值比较

图 5.7 展示的是不同方法中每次迭代后的 f1-score 变化曲线。由图可见，CEF

的曲线比 MS 的更平稳，可见 CEF 方法更加的健壮，且学习速率更快，同样证明了基于众包解释性反馈的主动学习优化方法能够得到更健壮更高效的分类模型。

5.4.5 不同 batch size k 对模型性能的影响

Batch size 即每次迭代挑选的样本数。除以上实验之外，我们对微博数据，（初始训练集为 20，待标注集总量固定为 200）。每次设置不同的 batch size，迭代到所有 200 条数量用完为止，比较不同 batch size 对模型最终效果的影响。

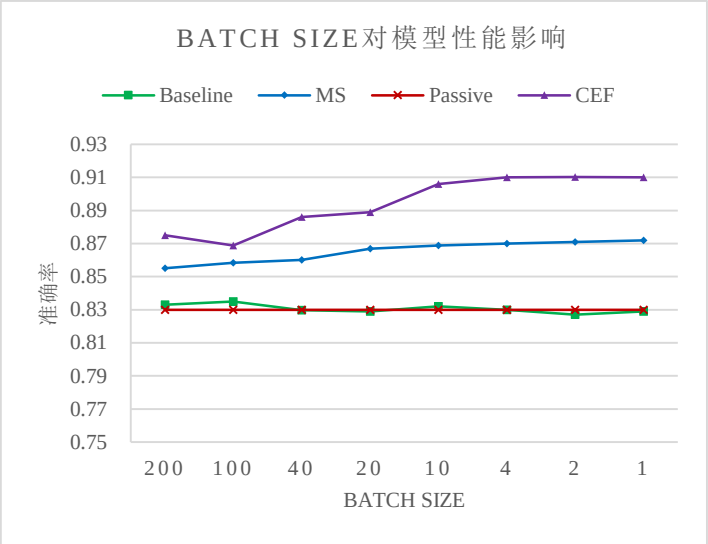


图 5.8 不同 batch size 对模型性能影响

图 5.8 中 Passive 是被动学习模型，即没有主动学习挑选样本的过程，将 220 条样本（20 初始训练集+200 标注集）一次性输入模型，只做一次模型训练，不做迭代，因此其准确率是一条直线。Baseline 方法的曲线几乎与 Passive 方法的曲线重合，因为每次随机挑选样本等同于一次性随机挑选样本，所以效果几乎保持一致，对分类结果没有明显提升。再来观察 MS 方法和 CEF 方法，分别有平均 3.5%和 6.4%的提升，并且曲线随着每次挑选样本数的减少逐渐增加，batch size=1 和 2 的时候两种方法的准确率能达到局部最高值。由此可以得出结论，batch size 越小最终的准确率越高，但是达到一定程度后曲线趋于平稳直线，提升不再明显，可能的原因是模型已经学习到了训练样本集所包含的几乎全部信息，由于训练集数量有限所以性能的提升效果也有限。

由以上 5 组实验能够看出，本文提出的引入解释性反馈的优化方法在不同实验环境下处理文本分类问题时，均能提升模型的分类性能。该方法能够融入更多人类潜意识信息从而更好地模拟人类对文本数据的认知行为。在实际应用中，样本的标记代价也是研究人员们重点关注的问题。接着，我们将对实验中的标注成本、时间代价和准确率进行综合讨论。

5.4.6 标注成本和准确率的讨论

如表 5.6 所示，显示了众包任务中每道题的单价：极性判断题（二选一）设置为 0.02 元/题；解释性反馈题（多项选择题）0.03 元/题。

表 5.6 众包任务单价	
极性判断题(元/题)	解释性反馈题(元/题)
0.02	0.03

接下来，我们讨论在迭代式环境下，不同方法在达到预设准确率 90%的情况下，所用的标注集个数 N_q、迭代次数以及花费的成本。如表 5.7 所示：

表 5.7 不同方法下的标注成本相关数据			
方法	达到 90%所用标注数量 N_q	迭代次数	标注成本（元）
Baseline	380	38	22.8
MS	280	28	**16.8**
CEF	140	**14**	**21**

由表 5.7 所示，当我们准确率要求设置为 90%时，本文提出的方法 CEF 所用标注数量和迭代次数最少，标注数量比 Baseline 和 MS 方法分别减少 63%和 50%，这意味着该方法能够减少大量的时间成本和标注成本，提高任务完成效率。另外，由于众包过程中获取了更多的理解性知识，不得不花费更多金钱，所以标注成本并没有比 MS 方法低，但依然少于 Baseline，节约了 7.9%的成本。本次实验由于客观条件的限制，并没有在大型数据集上进行测试，但可以预见的是，如果将我们的方法运用于大型数据集预测上，相信能够节省更多成本和时间上的开销，因此非常适合用于对实际问题的解决上。

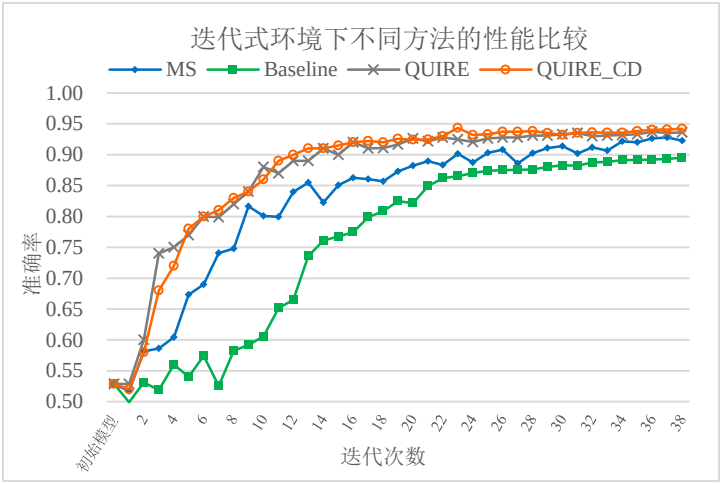
综合以上 6 组实验可得，本文提出的基于众包解释性反馈的主动学习模型优

化方法，能够在标注黄金训练集不充分的情况下，快速地提高模性效率，减少大量的标注数量以及时间成本，并且一定程度上减少标注成本，对解决实际问题较其他方法具有更大优势。

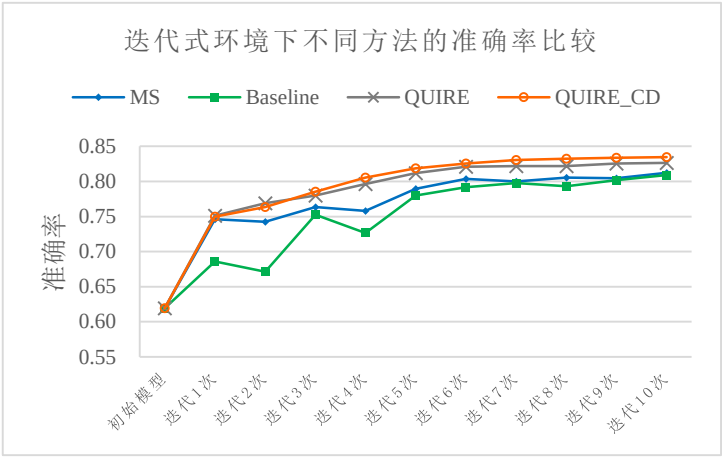
5.5 融入众包标注置信度的采样策略改进方法实验

5.5.1 不同挑选策略下的讨论

在迭代式环境下，比较不同挑选策略作用于两种评论数据集的最终分类效果，这里的实验设置和 5.4.4 中的实验设置一致，四种采样策略中的基础分类模型均选用 SVM 分类器。QUIRE_CD 方法中的参数 t 默认为 0.5。实验结果如图 5.9 所示。



(a) 微博评论数据集



(b) 酒店评论数据集

图 5.9 不同挑选策略下的准确率比较

由图 5.9 可见，本文提出的 QUIRE_CD 方法在两种数据集上，整体上都展示了较好的性能。QUIRE_CD 在最后一次迭代中比随机采样策略 Baseline 在两种数据集上分别提高了 4.2%和 2.5%，相比于 QUIRE 方法分别提高了 0.6%和 0.8%。通过比较本文的方法和 QUIRE 方法，我们发现，在酒店评论数据集中的提升效果比在微博评论数据集中略胜一筹，可能的原因是，酒店评论数据集比微博评论数据集在情感判断上更难一些，因此提升空间更大。

5.5.2 不同挑选策略下众包标注情况

接着，我们比较不同挑选策略下的众包标注准确率。为了保证实验的严谨性，我们选择相同的人群完成众包标注任务，并控制相同的标注环境（阿里众包平台）。

表 5.8 不同策略中的众包样本标签准确率

不同数据集（标注总数）	Baseline	MS	QUIRE	QUIRE_CD
微博评论数据（380）	0.96	0.95	0.95	**0.97**
酒店评论数据（100）	0.88	0.89	0.88	**0.93**

由表 5.8 可见，微博评论数据挑选的样本收集到的标注准确率普遍偏高均在 95%及以上，酒店评论样本相比较下则偏低。QUIRE_CD 的方法挑选出来的样本在阿里众包平台上标注准确率比其他三种方法的都要高，由此可见，该方法能够挑选出较为容易的样本，获取更高质量的标签。

5.5.3 不同影响因子下准确率的变化情况

由于本文提出的 QUIRE_CD 采样算法在酒店数据集上的提升效果更明显，我们在该数据集上，观察不同影响因子 t 对模型分类效果的影响。在一步环境下，挑选 6000 条均匀酒店评论数据集进行实验。实验中，初始训练集设置为总数据集的 0.5%，即 30 条；标注样本数量固定为 100 条。实验结果如图 5.10 所示。

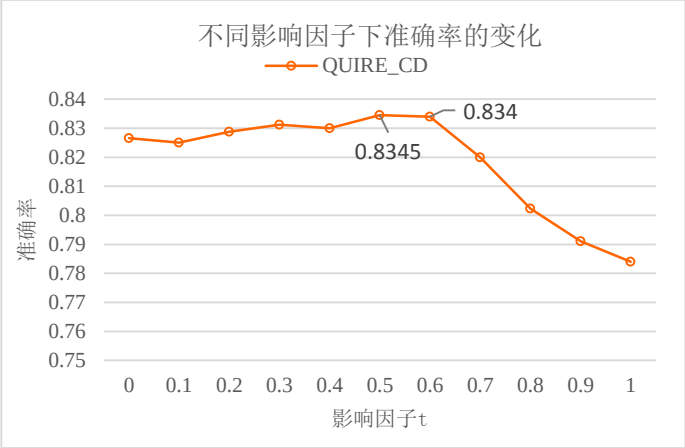


图 5.10 不同影响因子下准确率的变化曲线

由图 5.10 可见，模型准确率随着影响因子 t 的增加，先缓步增长然后从 0.6 开始大幅度下降。由此可以推断，在挑选过程中引入众包标注置信度，能够一定程度提升模型性能，但不能对其进行过分地约束，而忽略对样本信息量和代表性的控制，否则，分类效果会适得其反。

综合以上多组对照实验可以看出，QUIRE_CD 的采样策略能够得到更高质量的标签，并一定程度提升模型的性能上限，使得众包技术能够更好地融入主动学习模型中。

5.6 本章小结

本章节主要通过多组对照实验验证本文提出的两个方法在处理文本分类问题的有效性，其中具体介绍了实验的评价指标、实验设计方案、众包任务发布过程以及对实验结果的分析与比较。实验结果表明基于众包解释性反馈的优化方法使得模型学习速率比传统的方法更快，融入众包标注置信度的优化方法能够提升模型性能的上限，得到更优的分类器。

第六章 总结和展望

6.1 本文工作总结

本文首先介绍了基于众包的主动学习模型的研究背景和现实意义，其次讨论并总结了现有相关研究工作的不足，引出了本文的两个优化方法：从众包任务设计角度，提出一种基于众包解释性反馈的主动学习模型优化方法，从而获得更多人类对数据的潜在认识；从主动学习采样策略角度，保证样本充足的信息量和代表性的基础上融入众包标注置信度，挑选适合大众标注的样本进行标注，使得众包和主动学习模型能够更好地结合。最后，为了验证本文提出的方法的有效性，在两个文本分类数据集上进行多组对照实验。实验结果表明本文提出的两种基于众包的主动学习模型优化方法在处理文本分类问题上确实能够提升模型分类性能。

本文主要工作总结如下：

1. 从众包任务设计角度出发，设计一种新型众包任务，规则化解释性信息的形式，在收集数据标签的同时收集用户解释性反馈。本方法收集到的解释性信息能够挖据到数据潜在的重要特征，提升模型的学习能力。

2. 在现有的主动学习挑选策略 QUIRE 的基础上进行优化，提出一种基于众包反馈的主动学习挑选策略的优化方法，在采样过程中加入众包标注置信度的挑选标准，挑选适合人群的样本进行标注，从而降低标签噪音，优化模型输入。本方法能够将众包与主动学习模型更好地融合，并改善模型分类性能。

3. 通过多组对照实验，验证了本文提出的两种基于众包的主动学习模型优化方法在处理文本分类问题上的可行性和有效性。

6.2 未来工作

本文分别从众包任务设计角度和主动学习采样策略角度出发，提出两种基于众包的主动学习模型优化方法，从而挖掘更多用户对文本数据的潜在认识，并且

较好地融合众包和主动学习两种技术。然而，本文的方法仍然存在大量工作需要进一步完善与改进，接下来列出一些未来工作：

1. 本文的方法仅对文本分类领域的二个数据集进行实验并论证其有效性，之后的工作可以将该方法扩展到更多任务类型的数据上，例如图像标注数据集、实体识别数据集等。

2.本文主要通过提升数据特征权重的方式将解释性反馈融入分类模型中，之后可以考虑更复杂的方法来挖掘数据特征的重要性。

3.在融合众包标注置信度的优化方法中，本文给予参数赋予一个经验值，之后可以考虑提出一个有理论依据的计算方法并设计实验加以证明。

参考文献

[1] Andrew Mccallum K N. Employing EM in Pool-Based Active Learning for Text Classification[J]. Icml, 1998.

[2] Angluin D. Queries and concept learning[J]. Machine Learning, 1988, 2(4):319-342.

[3] Arasu A, Kaushik R. On active learning of record matching packages[C]// ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, Usa, June. DBLP, 2010:783-794.

[4] Atlas L, Cohn D, Ladner R, et al. Training connectionist networks with queries and selective sampling[C]// Advances in Neural Information Processing Systems. Morgan Kaufmann Publishers Inc. 1990:566-573.

[5] Balcan M F, Broder A, Zhang T. Margin Based Active Learning[C]// Learning Theory, Conference on Learning Theory, COLT 2007, San Diego, Ca, Usa, June 13-15, 2007, Proceedings. DBLP, 2007:35-50.

[6] Bellare K, Iyengar S, Parameswaran A G, et al. Active sampling for entity matching[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2012:1131-1139.

[7] Branson S, Horn G, Wah C, et al. The Ignorant Led by the Blind: A Hybrid Human–Machine Vision System for Fine-Grained Categorization[J]. International Journal of Computer Vision, 2014, 108(1-2):3-29.

[8] Cambria E, Nguyen T V, Cheng B, et al. GECKA3D: A 3D Game Engine for Commonsense Knowledge Acquisition[J]. 2016.

[9] Cohn D, Atlas L, Ladner R. Improving generalization with active learning[J]. Machine Learning, 1994, 15(2):201-221.

[10] David D. Lewis, William A. Gale. A sequential algorithm for training text classifiers[C]// Acm Sigir Forum. ACM, 1994:3-12.

[11] Dawid A P, Skene A M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1):20-28.

[12] Donmez P, Carbonell J G, Bennett P N. Dual Strategy Active Learning[C]// Machine Learning: Ecml 2007, European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007, Proceedings. DBLP, 2007:116-127.

[13] Garcia-Molina H, Joglekar M, Marcus A, et al. Challenges in Data Crowdsourcing[J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(4):901-911.

[14] Hodhod R, Huet M, Riedl M. Toward Generating 3D Games with the Help of Commonsense Knowledge and the Crowd[C]// Experimental AI In Games | An AIIDE 2014 Workshop. 2014.

[15] Hoi S C H, Jin R, Lyu M R. Large-scale text categorization by batch mode active learning[C]// International Conference on World Wide Web, WWW 2006, Edinburgh, Scotland, Uk, May. DBLP, 2006:633-642.

[16] Hoi S C H, Jin R, Zhu J, et al. Semi-supervised SVM batch mode active learning for image retrieval[C]// Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2015:1-7.

[17] Howe, Jeff. The Rise of Crowdsourcing[J]. 06 Jenkins H Convergence Culture Where Old & New Media Collide, 2006, 14(14):1-5.

[18] Huang S J, Jin R, Zhou Z H. Active learning by querying informative and representative examples[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2010:892-900.

[19] Ipeirotis P G, Provost F, Wang J. Quality management on Amazon Mechanical Turk[C]// ACM SIGKDD Workshop on Human Computation. ACM, 2010:64-67.

[20] Jia D, Krause J, Li F F. Fine-Grained Crowdsourcing for Fine-Grained Recognition[C]// Computer Vision and Pattern Recognition. IEEE, 2013:580-587.

[21] Kranjc J, Smailović J, Podpečan V, et al. Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform[J]. Information Processing & Management, 2015, 51(2):187-203.

[22] Leng Y, Xu X, Qi G. Combining active learning and semi-supervised learning to construct SVM classifier[J]. Knowledge-Based Systems, 2013, 44(1):121-131.

[23] Lewis D D, Catlett J. Heterogeneous Uncertainty Sampling for Supervised Learning[J]. Machine Learning Proceedings, 1994:148-156.

[24] Li C L, Ferng C S, Lin H T. Active learning with hinted support vector machine[J]. Journal of Machine Learning Research, 2012, 25:221-235.

[25] Li X, Guo Y. Active learning with multi-label SVM classification[C]// International Joint Conference on Artificial Intelligence. AAAI Press, 2013:1479-1485.

[26] Li X, Wang L, Sung E. Multilabel SVM active learning for image classification[C]// International Conference on Image Processing. IEEE, 2004:2207-2210 Vol. 4.

[27] Liu Y. Active learning with support vector machine applied to gene expression data for cancer classification.[J]. Journal of Chemical Information & Computer Sciences, 2004, 44(6):1936.

[28] Marcus A, Wu E, Karger D, et al. Human-powered sorts and joins[J]. Proceedings of the Vldb Endowment, 2011, 5(1):13-24.

[29] Mozafari B, Sarkar P, Franklin M J, et al. Active Learning for Crowd-Sourced Databases[J]. Computer Science, 2012.

[30] Mozafari B, Sarkar P, Franklin M, et al. Scaling up crowd-sourcing to very large datasets: a case for active learning[J]. Proceedings of the Vldb Endowment, 2014, 8(2):125-136.

[31] Muhammadi J, Rabiee H R, Hosseini A. A unified statistical framework for crowd labeling[M]. Springer-Verlag New York, Inc. 2015.

[32] Nguyen H T, Smeulders A. Active learning using pre-clustering[C]// International Conference on Machine Learning Icml. 2004:79.

[33] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[C]// International Conference on Language Resources and Evaluation, Lrec 2010, 17-23 May 2010, Valletta, Malta. DBLP, 2010.

[34] Sarawagi S, Bhamidipaty A. Interactive deduplication using active learning[C]// Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002:269-278.

[35] Settles B, Craven M, Ray S. Multiple-Instance Active Learning[C]// Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December. DBLP, 2008:1289--1296.

[36] Settles B, Craven M. An analysis of active learning strategies for sequence labeling tasks[C]// Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008:1070-1079.

[37] Settles B. Active Learning Literature Survey[J]. University of Wisconsinmadison, 2009, 39(2):127–131.

[38] Seung, H. S, Opper, et al. Query by committee[J]. Proc of the Fith Workshop on Computational Learning Theory, 1992, 284:287-294.

[39] Singla A, Tschiatschek S, Krause A. Actively Learning Hemimetrics with Applications to Eliciting User Preferences[J]. 2016.

[40] Thompson C A, Califf M E, Mooney R J. Active Learning for Natural Language Parsing and Information Extraction[C]// Sixteenth International Conference on Machine Learning. 1999:406--414.

[41] Tian T, Ning C and Jun Z. Learning Attributes from the Crowdsourced Relative Labels[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.

[42] Tong S. Support vector machine active learning for image retrieval[C]// ACM

International Conference on Multimedia. ACM, 2001:107-118.

[43] Tong, Simon, Koller, Daphne. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2001, 2(1):45-66.

[44] Vijayanarasimhan S, Grauman K. Cost-Sensitive Active Visual Category Learning[J]. International Journal of Computer Vision, 2011, 91(1):24-44.

[45] Wang J, Kraska T, Franklin M J, et al. CrowdER: crowdsourcing entity resolution[J]. Proceedings of the Vldb Endowment, 2012, 5(11):1483-1494.

[46] Xu Z, Yu K, Tresp V, et al. Representative Sampling for Text Classification Using Support Vector Machines[M]// Advances in Information Retrieval. Springer Berlin Heidelberg, 2003:11-11.

[47] Yan Y, Rosales R, Fung G, et al. Active Learning from Crowds[C]// International Conference on Machine Learning, ICML 2011, Bellevue, Washington, Usa, June 28 - July. DBLP, 2011:1161-1168.

[48] Zhang C, Chen T. An active learning framework for content-based information retrieval[J]. Multimedia IEEE Transactions on, 2002, 4(2):260-268.

[49] Zhao Y, Zhu Q. Evaluation on crowdsourcing research: Current status and future direction[J]. Information Systems Frontiers, 2014, 16(3):417-434.

[50] Zhong J, Tang K, Zhou Z H. Active learning from crowds with unsure option[C]// International Conference on Artificial Intelligence. AAAI Press, 2015:1061-1067.

[51] 冯剑红, 李国良, 冯建华. 众包技术研究综述[J]. 计算机学报, 2015(09):1713-1726.

[52] 刘康, 钱旭, 王自强. 主动学习算法综述[J]. 计算机工程与应用, 2012, 48(34):1-4.

[53] 孙欢. 众包标注的学习算法研究[D]. 浙江大学, 2015.

[54] 吴伟宁, 刘扬, 郭茂祖,等. 基于采样策略的主动学习算法研究进展[J]. 计算机研究与发展, 2012, 49(06):1162-1173.

[55]徐美香, 孙福明, 李豪杰. 基于主动学习的多标签图像在线分类算法[C]// 和谐人机环境联合学术会议. 2014.

[56]众包，智库百科，http://wiki.mbalib.com/wiki/众包

附录一 作者攻读硕士学位期间参与的科研项目与专利

[1] 媒体动态自组织关键技术研究与应用示范“数据驱动的媒体内容动态自组织及封装技术”，国家 863 项目。

[2] 专利“一种基于众包反馈和主动学习的文本分类模型优化方法”正式受理，申请号：201710205306.4。

致谢

时光荏苒，光阴如梭，随着两年半的研究生生活即将结束，也预示着要对二十多年的求学生涯说一声“再见”。回想起第一次来华师大的那场暴雨，似乎牵起了我和华师大的缘分，历历在目，仍似昨天。这几年在华师的日子，有欢笑，有泪水，无论是在学习和生活中，收获很多，遗憾也很多，在毕业之际，我由衷地感谢所有帮助、关心和陪伴我的人。

首先，我特别感谢我的指导老师——杨静老师。回首两年多的学习期间，杨老师无论是学习上还是生活中都给予我们无微不至的照顾。杨老师作为我们项目组的指导老师，带领我们解决一个又一个难题。感谢杨老师两年多来对我的学习和科研生活的指导，在她的悉心指导下我的论文才得以完成，在此表示我最真诚的感谢。

其次，我要感谢我的小伙伴和师兄师姐师弟师妹们。我觉得有这么一群同甘共苦的实验室小伙伴是我研究生阶段最大的收获，在这两年半时间里，我们一起学术、一起玩耍、互相勉励、共同进步。感谢实验室的师兄师姐们，在我遇到问题时，总会不耐其烦地给我讲解。感谢总会为实验室带来各种欢乐的师弟师妹们，正因为有了他们的存在，让我在专研学术和开发项目的每个日日夜夜里不再孤单。

最后，我要感谢我的家人和朋友，没有他们，就不可能有现在的我。在我 25 年的人生，18 年的求学生涯，是爸爸妈妈养育我、支持我、鼓励我，他们是我成长的不竭动力。我希望我能成为他们的支柱，他们的骄傲。希望他们能够永远幸福安康。

2017.9.21 于理科大楼 706