单位代码：10359　　密　级：公开
学　　号：2016110982　　分类号：TP181

合肥工業大學

Hefei University of Technology

硕士学位论文

MASTER'S DISSERTATION

论文题目：基于 Wikipedia 语料扩展的短文本数据流分类方法研究

学位类别：学历硕士

专业名称：计算机软件与理论

作者姓名：王海燕

导师姓名：胡学钢 教授

完成时间：2019 年 4 月

合　肥　工　业　大　学

学历硕士学位论文

基于 Wikipedia 语料扩展的短文本数据流分类方法研究

作者姓名：王海燕

指导教师：胡学钢 教授

学科专业：计算机软件与理论

研究方向：人工智能及其应用

2019 年 4 月

A Dissertation Submitted for the Degree of Master

Research on the Short Text Stream Classification based on the Corpuse Extension from Wikipedia

By

Wang Hai Yan

Hefei University of Technology

Hefei, Anhui, P.R.China

April, 2019

合 肥 工 业 大 学

本论文经答辩委员会全体委员审查，确认符合合肥工业大学学历硕士学位论文质量要求。

答辩委员会签名（工作单位、职称、姓名）

主席：专家工作单位，职称，姓名

委员：

导师：

学位论文独创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下进行独立研究工作所取得的成果。据我所知，除了文中特别加以标注和致谢的内容外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 合肥工业大学 或其他教育机构的学位或证书而使用过的材料。对本文成果做出贡献的个人和集体，本人已在论文中作了明确的说明，并表示谢意。

学位论文中表达的观点纯属作者本人观点，与合肥工业大学无关。

学位论文作者签名：　　签名日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解 合肥工业大学 有关保留、使用学位论文的规定，即：除保密期内的涉密学位论文外，学校有权保存并向国家有关部门或机构送交论文的复印件和电子光盘，允许论文被查阅或借阅。本人授权 合肥工业大学 可以将本学位论文的全部或部分内容编入有关数据库，允许采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：　　指导教师签名：

签名日期：　　年　　月　　日　　签名日期：　　年　　月　　日

论文作者毕业去向

工作单位：

联系电话：　　E-mail：

通讯地址：　　邮政编码：

致　谢

转眼间，研究生生涯即将结束，回顾过去的近一千个日日夜夜，有煎熬、有汗水也有快乐。在这近三年的研究生时间里，我收获到的不仅仅只有知识，还有对自己未来生活的规划，同时我的综合素质也得到了很大的提升。这些改变与老师们，与师兄师姐们的的谆谆教导和以身作则分不开。读研的这段时间是我往后的生活中值得珍藏和宝贝的美好回忆，在此我真挚的感谢那些学术道路上和做人原则上给予我指导和帮助的老师和同学们。

首先，我要感谢我的导师胡学钢教授和李培培副教授，他们既是我科研路上的导师也是我做人的指路明灯。感谢胡老师，是他将我领进了科研的大门，教会我如何学习他人论文，如何找到自己的研究方向，如何规划自己的学术生涯。李老师是一位对科研极度认真负责的人，她在我的论文方向和写作上给予了我最大的帮助，常常在我科研遇到瓶颈时给予我有用的建议。感谢这两位老师毫无保留的付出，是他们对待学术严谨的态度，对待生活积极乐观的心态影响着我。

其次，我要感谢数据挖掘“千人计划”团队中的吴信东、张玉红、吴共庆、李磊等各位老师，感谢他们在平时的主题报告会议上给予我的指导和帮助。同时，我还要特别感谢周鹏师兄、朱毅师兄、刘纳师姐、刘啸剑师兄、刘炉师兄、王勤勤师姐，在我的日常科研和生活中给予了我很多的帮助

另外，我还要感谢同实验室的杨帅、何路、吴咪咪、邵玉涵等同学，感谢同寝室的肖倩文、张陈、张晗，感谢他们在这三年的学习和生活路上的陪伴、帮助和支持。

我还要衷心的感谢我的家人，感谢他们在我读研这段时间的默默付出，是他们无私的支持和帮助，让我有坚持下去的勇气和决心，让我能够全身心的投入到科研的工作中。

最后，我要感谢评审我论文的专家们，感谢你们百忙之中抽出时间对我的论文的评阅和指点。

作者：王海燕

2019 年 4 月 10 日

摘　要

社交网络等领域产生了海量的短文本数据流，一方面，由于短文本自身长度短，语义信息不足，，流环境下的短文本数据具有的特征高维和概念漂移等特点，带来文本的高维稀疏问题，导致传统的文本分类方法难以直接应用。另一方面，随着短文本数据的快速产生，人工标注所有短文本数据不仅费时费力，且几乎是不可能完成，因此，如何在少量的有标记短文本数据的情况下，充分利用丰富的未标记短文本数据提升分类精度也是一大挑战。

针对以上问题，本文对短文本数据流分类算法进行了研究，其主要工作如下：

（1）概述已有的短文本分类的相关工作，包括：有监督短文本与短文本数据流分类方法以及半监督短文本与短文本数据流分类方法。

（2）针对短文本数据流存在的特征高维稀疏以及概念漂移问题，提出一种基于文本扩展和概念漂移检测的短文本数据流分类算法。该方法首先从 Wikipedia 获取外部语料用于扩展短文本，同时借助在线 BTM 模型选择代表性主题表示短文本，从而解决短文本的高维稀疏问题；其次，为检测短文本数据流中的概念漂移问题，提出一种基于主题的概念漂移检测算法；最后，该方法基于数据块构建集成模型，同时根据概念漂移检测结果利用当前数据块更新集成模型。实验结果表明：该方法在短文本数据流分类精度上表现优异，所提的概念漂移检测算法具有良好的检测性能。

（3）针对大量标记数据缺失问题，提出一种基于标签传播的半监督短文本数据流分类算法。首先，从 Wikipedia 中获取外部语料，并借助 Word2Vec 模型训练获得原始词向量集合用于短文本数据的向量化表示，以解决短文本数据流的特征高维稀疏问题。其次，分别针对标记和无标记数据构建分类器与聚类器形成集成模型，并采用基于簇相似度的方法传递聚类簇间的标签信息为其打上标签。同时，为了适应概念漂移，提出一种基于聚类簇的概念漂移检测机制。实验结果表明该方法能有效处理带缺失标签和概念漂移的短文本数据流分类问题。

关键词：短文本数据流；Wikipedia 外部语料库；在线 BTM；概念漂移；标签传播；

ABSTRACT

Many real-world applications such as Social Networks have produced huge-volume short text streams. SOn one hand, traditional text classification methods are difficult to be applied in the handling of short text stream, due to the short length of texts, the lack of semantic information, and the characteristics of high dimension and concept drifts, which causes the serious high-dimension and sparsity problem. On the other hand, with the rapid growth of short texts, it is not only time-consuming, but also almost impossible to manually label all short texts, thus, how to improve the classification accuracy by making full use of massive unlabeled short texts is also a large challenge in the short text stream classification with fewer labeled short texts.

In view of the above problems, this dissertation focuses on short text stream classification based on corpus extension from Wikipedia, and our main work is as follows:

(1) We summarize the relevant work of existing short text classification approaches, including the classification approaches of supervised and semi-supervised short text classification without/with data stream..

(2) Due to the characteristics of high dimensional and sparse features, and concept drift in the short text stream, we proposed a short text stream classification algorithm based on text extension and concept drift detection. Specifically, in the method, to make up for the sparsity of data, we firstly obtained the external corpus from Wikipedia to extended short text streams, and used online BTM to select representative topics instead of the word vector to represent the feature of short texts. Secondly, we proposed a concept drift detection method based on the topic model to detect the hidden concept drifts in short text streams. Thirdly, we built an ensemble model using several data chunks and updated with the newest data chunk and results of the concept drift detection. Experimental results showed that this method has excellent performance in short text stream classification, and the proposed concept drift detection algorithm had good detection performance.

(3) Due to the lack of labeled short texts and the massive unlabeled data, we proposed semi-supervised short text stream classification based on label propagation. Firstly, to solve the problem of high dimension and sparsity of features due to the short length of texts, the original word vector set of external corpus from Wikipedia was obtained by Word2vec, which was used to represent the feature space of short texts.

Secondly, the ensemble model was built using the classifiers and cluster models learnt from labeled and unlabeled data respectively, and then the cluster based similarity method was proposed for label propagation. To adapt to the concept drift, a new concept drift detection algorithm based on clusters was proposed. Experimental results showed that the proposed method was effective.

KEYWORDS： short text stream; concept drift; Wikipedia external corpus; online BTM; label propagation.

目　录

插图清单

表格清单

第一章 绪 论

实际应用领域如：社交网络平台产生了大量的短文本数据流，为了获取其中潜在的有价值的信息，针对短文本数据流的挖掘任务变得尤为重要。本章将主要阐述短文本数据流分类挖掘的研究背景、实际意义和本文的主要研究内容。

1.1 研究背景及意义

随着移动互联网和智能手机的普及，越来越多的人将网络纳入日常生活中不可或缺的一部分，人们通过网络观察整个世界，并发表自己的观点，例如根据新浪微博《2017 年微博用户发展报告》[1]，截止到 2017 年 9 月，微博的日活跃人数共 1.65 亿，与 2016 年同期相比增长 25%；月活跃用户达到 3.76 亿，较上一年度同期增长 27%。除此之外，像 Twitter、微信、QQ 等人们常用的社交软件，像淘宝、京东等的大型购物网站，每秒都有可能会产生数万条文本数据，由于这些文本数据具有长度短小、信息不足，内容稀疏，且产生速度快、数量大等特性，我们称之为短文本数据流。

短文本数据流蕴含着丰富的研究价值和商业价值，例如新浪微博作为一款有影响力的信息分享和交流平台，它为广大研究者、企业用户和政府部门提供了丰富的短文本数据[2]，这些海量的短文本数据流可以监听和分析网络舆情，帮助使用者快速了解研究对象的相关热门话题，发现未知的洞察，以及危险预警等。除此之外，我们可以利用相关的数据挖掘算法分析微博的短文本数据流，用以检测各个行业相关账号的表现和动态，分析比较分析格局和微博互动效果等。

为此，合理的分析与利用这些短文本数据，对广大研究者来说，有助于拓宽研究视角，激发学术激情，为我国学术研究提供更加丰富的成果；对商业用户来说，有助于了解客户群体的喜好，从而生产更适应客户喜好和行为的产品；对政府部门来说，有助于了解社会舆论走势，解决民生问题，从而建立一个为人民服务，对人民负责的好政府。

由于短文本的研究价值和其本身的特性，使得国内外研究者们越来越重视这方面的研究，由此出现了各种各样的关于短文本的研究，包括针对短文本的分类[6]、聚类[39-40]、主题发现[41-42]、特征选择[43-44]等数据挖掘算法。为了获取短文本数据中潜在的有价值的信息，有关短文本分类的研究变得尤为重要。因此，本文主要研究的内容是在流环境下的短文本分类，即短文本数据流分类。

1.2 研究问题和挑战

1.2.1 课题来源

本文的主要研究内容得到以下项目的资助：国家重点研发计划项目课题六：“基于数据融合的煤矿典型动力灾害多元信息挖掘分析技术”（No. 2016YFC0801406）；国家自然科学基金：“面向多源高维数据流的在线特征选择与分类方法研究”（No. 61673152）；国家自然科学基金：“多标记文本数据流分类方法研究”（No. 61503112）。

1.2.2 短文本数据流分类问题与挑战

与传统的长文本不同，短文本长度短，没有足够的语义信息，例如，一条推文被限制在 140 个字以内[27]，从而导致短文本数据在内容上的严重稀疏性。根据 2008 年 11 月由商务印书馆出版的《现代汉语常用词》，现代社会生活中较为稳定的，使用频率较高的汉语普通话常用词语有 56008 个。因此，如果采用传统的词袋模型表示中文短文本，就意味着表示每条短文本的向量中可能仅有几十维甚至几维有值，其余几万维都为零，从而造成短文本数据在表示形式上的严重稀疏性。

随着互联网的快速发展，短短的几万维向量肯定是不足以表示海量的短文本数据，可能需要几十万维甚至几千万维的向量来表示一条短文本，因此短文本的高维问题也是处理短文本数据亟需解决的问题之一。

由于短文本数据具有文本长度短，特征高维稀疏等问题，导致传统的文本分类方法很难有效处理，例如 SVM[3]，随机森林[4]，贝叶斯网络[5]等分类方法在处理长文本上具有很好的效果，但对于短文本本身所具有的特性就很难适应。另外，在流环境下的短文本数据还具有数据产生速度快、数量大，以及随着时间推移文本信息会发生潜在的漂移等现象，如：在线购物会随着季节的变化，个人的成长甚至商家的服务模式的变化而改变评论的话题；个人的微博也会随着心情的变换、个人兴趣的变化而改变每天的主题，我们称这种流环境下的潜在漂移现象为概念漂移。流环境下短文本所表现出的概念漂移现象也导致传统的批处理算法很难去快速有效的适应。

主题模型[45]通过挖掘文档的隐含主题构建模型，实现文档到词汇的映射。近年来，借助主题模型解决短文本分类成为常用的方法之一。本文第三章通过借助主题模型挖掘来自 Wikipedia 的外部语料库的隐藏主题用以扩展短文本，同时将扩展后的短文本表示成主题从而缓解短文本数据流的特征高维稀疏问题。另外为适应流环境下短文本数据出现的概念漂移现象，我们提出一种基于主题的概念漂移检测算法。

1.2.3 半监督短文本数据流分类问题与挑战

在实际应用中，由于标记短文本数据的缺少导致短文本数据流分类方法很难取得良好的成效。这是因为标记数据的获取往往需要消耗大量的人力物力，甚至需要依赖于少数领域专家来完成。例如在计算机辅助诊断中，我们可以很方便的拿到日常体检数据作为训练数据，但是对于日常体检的结果却是需要医学权威专家来提供的，而让医学专家来诊断所有的体检结果往往是不现实的。但是相比较而言无标记数据的获取就相对简单，且获取的数量巨大。因此，本文第四章主要解决实际应用中短文本数据流的类标签缺失问题，考虑如何有效利用丰富的无标记短文本数据和少量的标记短文本数据来共同提升所学模型的泛化能力。

目前，针对解决标记数据缺少，未标记数据丰富问题有三大主流技术，分别是直推式学习[55-56]、主动学习[57-58]以及半监督学习[28-38]。所谓直推式学习是假设未标记数据就是最终要用来测试的数据，学习的目的就是在这些数据上取得最佳泛化能力，而主动学习是通过某种算法查询到最有用的未标记样本，并将其交给专家进行标记，然后标记的样本用于训练分类模型来提高模型的精确度。不同于直推式学习仅关注在未标记数据上的预测性能以及主动学习依赖于人工干预不同，基于半监督的学习通过自动利用标记数据和未标记数据，学习获得整个数据分布上具有强泛化能力的模型。本文第四章借助半监督学习的思想来解决实际应用中短文本数据的类标签缺失问题。

考虑到无标记数据不能直接构造分类器进行分类，本文第四章采用聚类模型将相似的无标记数据聚类获得聚类簇，然后通过已有的少量标记数据向每个聚类簇进行标签传播，最后将有标记的聚类模型联合标记数据构建的分类器构成集成模型分类新的短文本数据从而提升准确率。

1.3 本文组织结构

本节主要介绍了全文的组织结构。本文的内容共分为五章，各章的主要内容和安排如下：

第一章：绪论，首先对本文的研究背景与意义做了简要介绍。其次，介绍了有关短文本数据流分类和半监督短文本数据流分类的研究问题以及面临的挑战。最后，给出了本文的整体组织结构。

第二章：相关工作概述，首先介绍了有监督短文本分类和有监督数据流分类的相关研究，然后给出了半监督短文本分类的相关研究，最后由于流环境下的半监督短文本分类研究太少，基于半监督的数据流分类研究被给出。

第三章：针对短文本的稀疏性问题和流环境下的概念漂移问题，提出了

一种基于文本扩展和概念漂移检测的短文本数据流分类算法。实验结果表明该分类算法能够有效的适应存在概念漂移的短文本数据流。

第四章：标记短文本数据的获取通常是费时费力、且获取的数量有限，本章提出了一种基于标签传播的半监督短文本数据流分类算法用以解决标记数据稀少问题。

第五章：总结与展望，对本文的研究内容进行总结，并进一步分析现阶段工作中存在的问题和不足，给出了今后研究的方向。

1.4 本章小结

随着移动互联网和智能手机的快速普及，类似于微博、Twitter 这样的社交网络平台，每天都会产生大量的短文本数据流，这些短文本数据流蕴含着丰富的商业价值与研究价值，有助于政府、企业决策以及个人学习。因此，为了获取短文本数据流中潜在的有价值的信息，短文本数据流分类变得尤为重要。本章作为绪论，首先简要介绍了短文本数据流的研究背景与意义，然后介绍了在短文本数据流分类问题的研究过程中遇到的挑战，最后给出了本文的组织结构。

第二章　相关工作概述

本章将概述有监督短文本分类和有监督数据流分类的研究现状，进而针对标记数据缺失问题，总结半监督短文本分类和半监督数据流分类的相关工作。

2.1 引言

随着互联网技术的迅猛发展，短文本作为一种新形式文本广泛存在于各种应用中，例如微博、聊天记录、用户评论、新闻标题等等。目前，针对短文本学术上没有给出严格的定义，我们普遍认为短文本是由几个或者几十个词组成，或者由几句话构成[6]。由于文本短小，没有足够的语义信息，导致了短文本数据承担着严重的稀疏性问题，因此传统的文本分类方法[3-5]很难直接应用到短文本分类上。

为解决短文本存在的稀疏性问题，提升文本分类的准确率，国内外研究者提出各种方法扩展短文本，但随着互联网用户的增加，短文本数据出现暴增，例如各大社交网站、购物网站每时每刻都会涌现出数以万计的评论，短文本数据已经呈现出一种流式的状态，我们称之为短文本数据流，该短文本数据流除了具备短文本的稀疏性问题外，其文本数据流存在的特征高维问题，由于数据连续不断的到来所带来处理的及时性问题、数据流随时间会产生潜在的漂移等问题都需要被解决。除此之外，在实际应用中短文本数据的到来是源源不断的，人工标注这些短文本数据不仅费时费力而且几乎是不可能完成的任务，但丰富的未标记数据同样具有丰富的语义信息，因此在面对少量的标记数据和丰富的无标记数据前提下，如何提升短文本分类的效果变得尤为重要。

本章除了详细介绍国内外研究者们在有监督短文本分类上做出的相关成果外，还会介绍目前关于有监督短文本数据流分类的相关研究。另外针对实际应用中的类标签缺失问题，本文主要研究在半监督学习框架下如何提升短文本数据流的分类效果，因此本章后续将介绍半监督短文本分类以及半监督数据流分类上的相关研究。

2.2 有监督的短文本分类方法

目前，针对有监督短文本分类的研究主要集中于解决短文本的稀疏性问题，外部语料库具有充足的语义信息，借助外部语料库扩展短文本是一种常见缓解短文本稀疏性的方法，常见的方法有基于搜索引擎扩展短文本的分类方法和基于主题模型扩展短文本的分类。但是外部语料库的获取必须要和所分类的短文本相关，其获取的质量直接影响到短文本分类的效果。为了充分利用短文本自身的信息，

基于自身的隐藏规则和统计信息的短文本分类也受到研究者的关注。另外，随着深度学习在计算机视觉、语音识别以及自然语言处理上的所取得巨大发展，作为自然语言处理中的一个分支，基于深度学习相关技术的短文本分类也成为重点研究对象。

2.2.1 基于搜索引擎的短文本分类方法

为了缓解短文本的稀疏性，研究者们通过将短文本视为一条查询语句，借助搜索引擎查询相关信息从而扩展短文本。Bollegala 等[7]提出了一种借助 Web 搜索引擎检索到的两个词的页面数和文本片段来估计语义相似性的方法。Wang 等[8]基于伪相关反馈理论，提出了一种新的基于 Web 搜索引擎的短文本扩展方法。在保持语义信息不变的情况下，首先将短文本作为查询语句放入 Web 搜索引擎中搜索获取相似的语料信息，然后从获取的语料信息中抽取特征向量扩展短文本，最后借助传统的分类方法分类短文本。通过模拟人在互联网上搜索信息的过程，Tang 等[8]提出一种端到端的学习方法 ExpaNet 用于短文本扩展，如图 2.1 所示。给定一条短文本，首先在检索模块中借助搜索引擎从外部语料库中检索一组可能相关的文档，然后短文本和获取的相关文档均用词向量进行表示，接着将向量表示后的短文本和相关文档丢入注意力机制（soft/hard attention）中获得检索记忆，该注意力机制被用于确定哪些文档是值得研究的，最后借助 GRU 模型根据注意机制得到的结果整合原始短文本。为了能多次迭代扩展短文本，被整合后的文本可以继续作为一个查询语句重复上述步骤，通过这种迭代扩展短文本的方式丰富短文本的语义信息，缓解稀疏性问题。

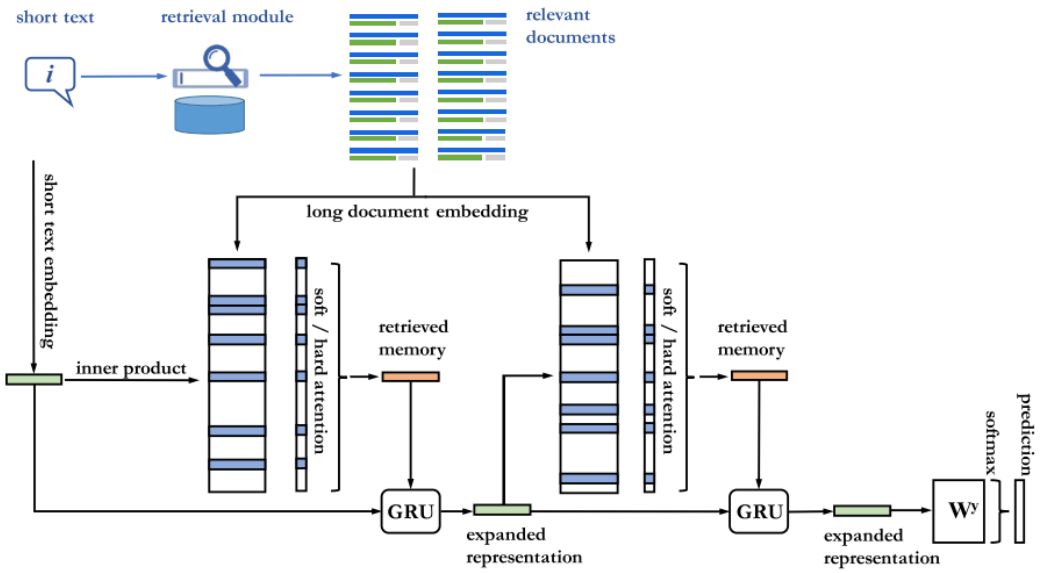

图 2.1 ExpaNet 模型结构

Figure 2.1 ExpaNet model structure

上述基于搜索引擎的短文本分类的相关研究借助搜索引擎从网上数据或者外部语料库中获取相关信息扩展短文本，从而丰富了短文本的语义信息，缓解了短文本的稀疏性问题。但是其搜索引擎的选择严重影响到短文本扩展的质量，另外从海量的网上数据或者大型外部语料库中获取相关信息也是非常耗时的。

2.2.2 基于主题模型的短文本分类方法

主题模型[41,45]能够挖掘数据集中潜在的语义信息，基于主题模型的短文本分类方法成为目前解决短文本稀疏性的研究热点之一。Phan 等[10]提出了一个基于隐含主题的框架用于扩展短文本，其基本思想是先借助 LDA 主题模型[45]从外部语料库中挖掘隐含主题，然后根据构建的主题模型推断短文本的主题分布，最后选择概率高的主题作为词扩展到短文本中，从而丰富短文本的语义信息，使得短文本更具主题导向性。Bouaziz 等[11]借助 LDA 主题模型从单词和文本两个层面上扩展短文本，同时为减少随机特征的选择，利用特征间的语义关系构建语义随机森林。同样借助 LDA 主题模型，Vo 和 Ock[12]从多种语料库中挖掘隐含主题，考虑短文本中其他词的语义关系寻找最合适主题扩展短文本。Zhang 和 Zhong[13]提出了一种新的短文本分类框架，该框架中词对应的主题会被视为新的词整合到短文本中，从而扩充短文本的语义信息。Chen 等[14]发现基于词袋模型的表示方法会导致某些不具备相同术语的短文本很难被正确分类，因此他们提出一种基于 LDA 主题模型和 K 近邻来提升短文本分类方法。如果两个短文本中某些术语具有相同的几个隐藏主题，则考虑文本之间相似，然后借助 K 近邻方法进行分类。

上述借助 LDA 主题模型扩展短文本在一定程度上缓解了短文本数据的稀疏性问题，但是由于 LDA 本身是依赖词袋模型的假设，忽略了短文本数据中的词序和短语，因此基于 LDA 的短文本扩展也受到了限制。主题的 N-gram 模型[59]（TNG）是基于上下文语义发现主题和短语的模型，Sun 等[15]借助 TNG 模型构建一个特征扩展库，同时提出一种短文本的主题权值向量计算方法获得短文本的主题倾向，然后根据主题倾向从特征扩展库中选择合适的词或者短语扩展短文本。

2.2.3 基于隐藏规则和统计信息的短文本分类方法

为了有效分类短文本数据，Kim 等[16]提出一种基于语义标注的语言独立核方法（LIS）。LIS 核方法被设计在不需要语法标签和词汇数据库的情况下，有效的分类短文本数据。短文本的稀疏性问题严重影响分类结果，Gao 等[17]引入了结构化稀疏表示用于短文本分类，同时为了提高分类的有效性，他们提出了一种凸包顶点选择方法，能够有效减少数据相关性和冗余性。Zhang 等[18]研究基于词汇关联规则的短文本分类方法，首先，挖掘训练集中存在的强关联规则，然后将这些强关联规则加入到短文本特征中用以提高短文本的特征密度，从而提升短文本分类的

准确性。Rao 等[19]为了缓解短文本稀疏性，提出了一种主题级最大熵模型（TME）用于短文本上的社交情感分类。TME 模型通过挖掘短隐含主题、多重情感标签以及众多读者的共同评分来生成主题级特征，通过将特征映射到概念空间还能解决最大熵原理中的过拟合问题。

上述基于隐藏规则和统计信息的短文本分类方法，虽然充分挖掘了短文本数据自身的隐藏规则和统计信息，但短文本本身存在的语义信息不足的问题并没有得到解决，因此其分类效果很难大幅度提升。

2.2.4 基于深度学习相关技术的短文本分类方法

近年来，深度学习的发展有效的推动了自然语言处理相关工作的进步。Li 等[20]提出了一种基于实体知识库和主题增强词嵌入（TEWE）获取短文本特征进行推文的主题分类，这种方法称为 TweetSift。基于实体知识库产生的特征是通过短文本本身和所属用户两个角度获得，TEWE 试图利用隐藏主题模型为每个短文本中的词分配主题，然后基于主题和词训练词向量，以便在不同语境下的相同的词有不同的向量值。TweetSift 通过整合基于实体知识库和 TEWE 产生的特征进行分类。Wang 等[21]提出了基于词嵌入聚类和卷积神经网络扩展短文本从而缓解数据稀疏性和语义敏感性问题。为了丰富短文本的语义信息，他们引入外部语料库进行词嵌入预训练从而初始化查找表，同时提出一种基于密度峰搜索的聚类算法用于短文本扩展。但由于词向量中词的数量巨大，上述聚类算法会消耗大量时间，因此该算法很难直接应用到实际中。为了加快聚类时间，Sotthisopha 等[22]提出了一种快速语义扩展的多通道卷积神经网络用于短文本分类。该方法可以减少计算时间和资源的需求，首先，微型批处理 Kmeans++算法被用于聚类词向量，其次，Jaro-Winkler 相似性被用于从已有词向量中寻找与没有词向量的词的最相似的词用以增加词在词向量上的覆盖率，最后借助多通道的 CNN 整合原始特征和扩展特征进而分类短文本。另外，由于有限的标记数据，Yan 等[23]将短文本将短文本分类视为小样本学习任务，提出一种基于孪生卷积神经网络和小样本学习的短文本分类框架。

主题模型旨在挖掘词在文章中的共现关系，腾讯 AI 实验室与香港中文大学合作，联合提出了一个主题记忆网络用于短文本分类[24]，该主题记忆网络可以分为三个部分：神经主题模型、主题记忆机制以及文本分类器。其中神经主题模型主要用于学习主题表示，主题记忆机制主要用于将学到的主题表示映射到对文本分类有用的特征空间当中，文本分类器主要用于预测短文本的标签。该方法除了在短文本上取得不错的分类效果，还能挖掘连贯的主题。

2.3 有监督的短文本数据流分类方法

目前针对短文本数据流分类的研究工作还很少。比较具有代表性的工作有，Bouaziz 等[25]提出 IGLM 模型，通过不断更新分类器提高数据流分类。首先，根据初始训练集训练随机森林分类器，其次，当有数据到来时先利用初始分类器进行分类，同时结合主动学习的方法将错分类的短文本加到训练集中，通过计算先前训练集与当前错分类短文本信息增益的差值决定是否更新分类模型。Ren 等[26]提出分层多标签短文本数据流分类。首先，基于实体链接和查询语句的排序方法扩展短文本，然后，通过将主题分成动态全局主题和局部主题构建动态概率主题分布，最后，使用基于块的结构优化策略分类短文本。Li 等[27]提出一种增量式的集成模型适应短文本数据流，首先，公开语义网络 Probase 被用于扩展特征空间，主要通过引入更多的基于短文本隐藏术语的语义上下文信息来弥补数据的稀疏性，同时为减少噪音影响，基于语义信息消除所有术语的歧义。然后基于概念簇的主题漂移检测算法被提出用于追踪数据流中的主题漂移。最后通过构建增量式的集成模型来预测短文本数据流。

2.4 半监督的短文本分类方法

上述有监督短文本分类算法虽然取得了良好的分类效果，但是为保证分类器具有良好的泛化性能，这类方法需要大量有标记的短文本数据作为前提。有标记短文本数据的获取通常是困难的，因为对这些数据进行标注往往需要有关专家的参与，例如对新闻评论数据的标注可能就需要有新闻工作经验的人的参与，不仅如此，通过这种方式标注短文本获取的数量也是有限的。但是无标记短文本数据的获取就相对简单的多，且获取的数据量也非常可观，因此基于半监督短文本分类方法也越来越受到研究者们的关注。

目前已有的基于半监督短文本分类的研究工作相对较少，主要可以简单分为三类：基于生成式模型方法[28]，基于自学习与协同训练的方法[29-32]以及基于图的方法[33,60]。

基于生成式模型的半监督文本分类方法简单、直观，通过少量标记数据和丰富未标记数据估计模型参数，具有良好的泛化能力。为解决标签瓶颈问题，Cai 等[28]提出基于随机子空间的最大期望半监督短文本分类方法（RS-EM）。RS-EM 采用迭代的方法进行 EM 训练，首先基于随机子空间方法从整个特征空间获得多个子空间，然后在每个子空间上借助朴素贝叶斯技术训练一个分类器从而构成集成分类器，最后在每一轮 EM 迭代时集成模型用于扩大标记数据集。RS-EM 方法通过将随机判别理论和半监督 EM 算法结合来弥补标准 EM 算法的过度训练问题。

基于自学习与协同训练方法研究较多，Chan 等[29]提出一种新的桥接方法作为基分类器用以提升自学习与协同训练等传统的半监督算法，不同于现有的将未标记数据直接转化为标记数据的半监督分类方法，该方法中未标记数据被用于为标

记数据和被分配标签的数据之间提供链接；Yin 等[30]采用半监督学习和支持向量机对传统方法进行改进；Silva 等[31]提出了一种半监督学习框架，该框架在未标记数据上构建相似矩阵用以捕获未标记数据丰富的信息，并将获得的相似矩阵和标记数据构建的分类器相结合用于半监督学习；Li 等[32]提出基于融合相似性测量方法的短文本半监督分类算法。以上的几种方法虽然利用丰富的未标记短文本数据和有限的标记短文本数据来提升分类效果，但忽略了短文本自身的稀疏性问题。

另外，基于图的半监督短文本分类方法有，Widmann 等[33]为了缓解短文本的稀疏性，提出基于图的半监督短文本分类方法，短文本和特征被表示成结点，结点之间的权值是基于词序、内容相似性以及词频获得。张倩和刘怀亮[60]提出了一种基于图结构的半监督短文本分类方法，该方法首先通过对少量标记短文本和海量未标记短文本进行基于图结构的文本表示，然后利用稀少的标记短文本构建初始分类器从而预测未标记的短文本数据，通过不断迭代的方式实现未标记短文本向标记短文本转化，最后更新扩大标记短文本数据集从而训练最终的分类器。基于图表示短文本解决了基于词袋模型表示短文本的稀疏性问题，从而提升短文本的分类效果。

但是，随着互联网的发展，短文本数据呈现出流的形式，以上的批量式半监督短文本分类方法已不再适用呈现特征高维稀疏、概念漂移等特性的短文本数据流。

2.5 半监督的数据流分类方法

已有基于半监督的数据流分类方法相对比较少，主要有 Zhang 等[34]提出一种基于一致性权值设置的集成模型，通过计算簇中心距离向聚类器传播标签信息，然后结合标记数据和未标记数据分别训练分类器和聚类器共同挖掘数据流。为解决不完全标记数据流和重复概念漂移现象，Li 等[35]提出了一种基于增量式单决策树模型的半监督数据流分类算法，被称之为 REDLLA，REDLLA 采用基于 Kmeans 的聚类算法在决策树的叶子结点上产生概念簇，同时借助不同时间下概念簇之间的偏差检测反复出现的概念漂移。同样为了解决数据流的重复概念漂移问题，Feng 等[36]提出一种表示学习的增量式半监督分类方法。Zhu 等[37]为学习标记和未标记的流数据，提出一种基于在线最大流算法的半监督学习系统。考虑从标记数据和无标记数据中学习图并动态的更新该图，用以适应在线数据的添加和收回，其中样本表示结点，边由结点之间的相似度表示。在从生成的非平稳图中学习时，他们通过增加和消除路径去更新最大流并从理论上保证了更新的最大流等于重新训练的最大流。为了分类样本，他们在当前最大流上计算最小割以便于最小化相似样本对的数量，从而将样本分类到不同的类别中。Hosseini 等[38]提出基于簇的分类器集成方法用以适应非平稳数据流的半监督分类算法。由于实际短文本数据流的

特征高维稀疏等特性，导致以上的基于半监督的数据流分类方法很难取得良好的分类效果。

2.7 本章小结

本章首先介绍了针对短文本的稀疏性问题，国内外研究者借助外部语料库扩展短文本从而提升短文本分类效果的相关研究。由于短文本的文本长度短小，将短文本视为一条查询语句，借助搜索引擎搜索相关信息扩展短文本是常用短文本扩展方法。但该方法依赖于搜索引擎的质量，如果搜索的结果不合适，其短文本扩展的质量也会很差。主题模型能够挖掘数据集隐含的语义信息，因此，借助主题模型挖掘外部语料库的隐含主题信息，从而扩展短文本也是研究的热点之一。但基于主题模型的短文本扩展需要预先定义主题数，如果预先定义了不合适的主题数，其扩展的效果也会受到影响。另外，上述两种扩展方法需要选择合适的外部语料库，如果选择的外部语料库与需要分类的短文本数据相关性不够或者完全不相关，其扩展后短文本的分类效果可能很难得到提升甚至有所下降。通过充分挖掘短文本的隐藏规则和统计信息提升短文本分类效果也吸引了很多的研究者，虽然提升了分类效果，但由于短文本本身的稀疏性问题并没有得到解决，因此效果有限。另外，随着深度学习在自然语言处理上的广泛应用，基于深度学习相关技术的短文本分类研究逐渐成为了目前的研究热点，例如近段时间提出的主题记忆网络，它是主题模型与文本分类在神经网络框架下的一次结合。

其次本章介绍了在数据流环境下的短文本分类算法的相关研究。随着互联网用户的增加，网络上每时每刻都会有数以亿计的短文本数据，因此标记全部的短文本数据是不现实的，且标记的成本也非常高。本章整理了基于半监督的短文本分类算法的相关研究，介绍了先前的研究者们是如何在少量标记数据和丰富的无标记数据的环境下提升短文本分类效果。本章的最后介绍了流环境下的半监督分类算法的相关研究。

第三章　基于文本扩展和漂移检测的短文本数据流分类算法研究

传统的短文本分类方法主要考虑如何解决短文本的高维稀疏性问题，然而短文本数据流除具有高维稀疏性外，还具有海量、快速以及随时间推移发生潜在的概念漂移等特点。为此，本章提出一种基于主题模型的文本扩展和概念漂移检测的短文本数据流分类方法。

3.1 引言

随着互联网技术和即时通信的飞速发展，网络用户和网络服务器每时每刻都会产生大量的短文本数据，包括微博、网上评论、即时信息等。这些不断产生的短文本数据被称为短文本数据流。短文本数据流在用户意图理解、问答系统以及智能信息检索等方面扮演着极其重要的角色。短文本数据流具有以下三种特性：每条短文本长度短且没有足够的语义信息，短文本数据具有严重的高维稀疏性问题；短文本的产生速度快，规模庞大；短文本数据流会随时间推移文本主题发生潜在漂移。基于这三个特点，很难直接在短文本数据流上应用传统的分类方法。

目前已有的针对短文本分类的研究[7-21]主要集中在处理短文本的稀疏性问题，主要分为两种方法，一是借助外部语料库扩展短文本用以提高分类效果，例如 Phan 等[10]、Bouaziz 等[11]以及 Vo 等[12]借助 LDA[45]主题模型，提出从外部语料库中挖掘隐藏主题用以扩展短文本。另一种是借助短文本数据本身潜在的规则和统计信息扩展短文本进而提升准确度，例如 Kim 等[16]提出 LIS（Language independent semantic）核用于短文本分类。上述短文本分类算法均属于批处理算法，但像新浪微博、Twitter 以及 Facebook 等社交平台每时每刻都会产生海量的短文本数据流，由于数据量巨大，极易造成维度灾难，且会随时间推移发生潜在漂移现象，因此批处理算法很难满足数据流的分类需求，我们急需一种专门针对不断变化的短文本数据流的分类算法。

为了分类短文本数据流，在线主题模型 OnlineBTM[41]（Online Biterm topic model）被提出。OnlineBTM 模型根据等时间片原则将短文本数据流划分为数据块，而每个数据块则是根据词对共现原则构建模型用以发现隐藏主题，这里我们用 biterm 表示出现在短文本中的无序词对。因此，与借助词共现的主题模型相比，OnlineBTM 更容易发现隐藏主题。但是，在短文本数据流中词对要比词更加稀疏，与此同时，OnlineBTM 并没有考虑到短文本数据流中的概念漂移问题。

针对上述 OnlineBTM 模型在处理短文本数据流时存在的稀疏性和概念漂移问题，本章提出了一种基于 OnlineBTM 模型的短文本数据流分类方法。首先，为了缓解短文本数据流存在的稀疏性问题，本章通过分析实验中用到的短文本数据流

从 Wikipedia 中获取相同主题的数据作为外部语料库用于扩展这些短文本数据；然后，为了表示扩展后的短文本数据流，本章借助 OnlineBTM 选择有代表性的主题而非词来表示。其次，为了有效检测并适应短文本数据流中的概念漂移问题，本章根据类标签将用于构建集成模型的数据块划分为类簇，然后计算新的数据块与每个类簇的语义距离，选择最小语义距离值判断新数据块相对于用于构建集成模型的数据块是否发生概念漂移。最后，根据是否发生概念漂移更新集成模型。

3.2 算法描述

本节详细介绍了基于文本扩展和概念漂移检测的短文本数据流分类方法，首先给出了问题定义和算法的整体框架；接着详细介绍了算法框架的技术细节，包括：基于短文本扩展的 OnlineBTM 模型、概念漂移检测以及集成模型的构建与更新。

3.2.1 问题定义

假定一个短文本数据流是由 N 个数据块组成，即 $D=\{D_i\}_{i=1}^{N}(N\to\infty)$，每一个数据块由一组短文本构成，即 $D_i=\{d_j\}_{j=1}^{|D_i|}$，其中 $|D_i|$ 表示第 i 个数据块（$1\le i\le N$）中短文本的总数。通常情况下，文本可以表示成一组向量，即 $d_j=\{(R^M,y_j)\,|\,y_j\in Y\}$，其中 R 表示文本空间的域，M 表示属性维度，Y 表示一组类标签。鉴于词袋模型表示短文本数据流会导致特征的高维稀疏问题，不利于分类表现。为了解决上述情况，本文借助来自 Wikipedia 的外部语料库扩展短文本数据流用以缓解短文本数据流的稀疏性问题，被扩展后的短文本数据块表示为 $D_i'=\{d_j'\}_{j=1}^{|D_i|}$，然后 OnlineBTM 主题模型被用于将扩展后的短文本数据块表示为 K 维的主题，即 $D_i''=\{d_j''\}_{j=1}^{|D_i|}$，其中 $d_j''=\{z_{j,k}\}_{k=1}^{K}$，$z_{j,k}$ 表示在 j^{th} 短文本中 k^{th} 主题值。最后，我们的目标是构建一个动态的分类器 $f:F_{\sum_{D_i''}}\to Y$，用以适应未知的短文本数据流以及发现短文本数据流中潜藏的概念漂移现象。

图 3.1 给出了所提算法的框架图，为了有效的处理短文本数据流，我们选择 H 个数据块 $S=\{D_1'',D_2'',\cdots,D_H''\}$ 构建 H 个基分类器从而建立一个集成模型 $E=\{f_1,f_2,\cdots,f_H\}$。当新的短文本 d_j 到来时，先经过扩展和主题表示得到 d_j''，集成模型通过公式 3.1 预测短文本 d_j''。

$$y^{*}=\arg\max_{y\in Y}\left(P(y,d)\right)=\sum_{h=1}^{H}w_{h,j}\,P_h(y,d) \tag{3.1}$$

其中 $w_{h,j}$ 表示使用基分类器 f_h 预测短文本 d_j'' 的权值。

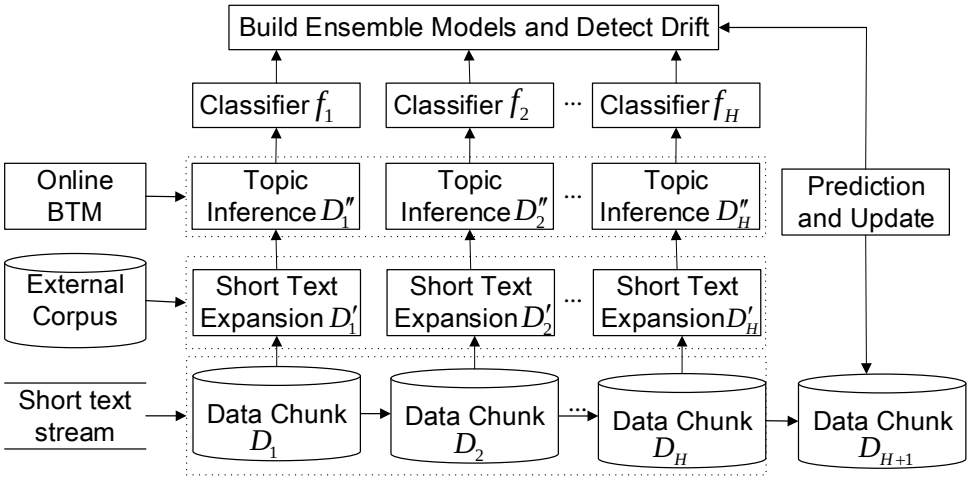

图 3.1 短文本数据流分类框架图

Fig 3.1 Framework of short text stream classification.

3.2.2 基于短文本扩展的 OnlineBTM 模型

BTM（Biterm Topic Model）用于短文本的主题发现，进一步被扩展用于短文本数据流上的主题发现，被称为 OnlineBTM[41]。与传统的主题模型相比，它通过词对共现模式来挖掘词之间的关系，加强了语义信息，缓解短文本的稀疏性问题。但是，在短文本数据流中词对要比词更加稀疏，因此我们提出基于短文本扩展的 OnlineBTM 模型，通过外部语料库扩展短文本以缓解短文本数据流存在的稀疏性问题，然后再经过 OnlineBTM 模型将短文本数据流表示成主题向量来解决短文本数据流的特征高维问题。

在介绍短文本的扩展和主题表示之前，我们首先介绍如何从 Wikipedia 中获取外部语料库，因为外部语料库获取的好坏直接影响到短文本扩展质量的好坏。

3.2.2.1 外部语料库的获取

外部语料库的好坏直接影响到短文本扩展的结果，因此，外部语料库必须满足两个特性[10]：一是外部语料库必须足够大内容足够全面，能够覆盖被分类的所有短文本；另一个是外部语料库必须与被分类的短文本数据流具有一致的主题。这里我们从 Wikipedia[1]中获取相关数据。具体获取方式如下：

首先我们从被分类的短文本数据流中提取关键词，其中每个类提取 50 个关键词；然后我们根据关键词借助 JwikiDocs 工具[2]从 Wikipedia 中获取相关的原始数据，其中每个关键词爬取 100 篇文章，通过上述方法我们总共获得 60600 篇原始文档，共计 1.34GB；最后我们对这些原始文档进行预处理，包括删除重复文档、HTML

[1] https://www.wikipedia.org/

[2] http://jwebpro.sourceforge.

标记和网页链接，以及去除停止词，我们得到 20968 篇文档，共有 268MB，有关获取外部语料库的详细数据表 3.1

表 3.1 外部语料库详情

Table 3.1: Details of obtained external corpus

Statistics of the crawled Wikipedia data
Raw documents: 1.34G; |documents| =60,600
Preprocessing: deleting duplicate pages, HTML tags and page links, removing stop words
Final documents： 268MB; |documents| =20,968; |vocabulary|=486,653; |total − words|=34,217,666

3.2.2.2 短文本扩展

在短文本扩展上，我们选择 LDA 而非适合短文本的 BTM 挖掘外部语料库中的隐含信息。主要原因在于上述从 Wikipedia 中获取的外部语料库属于长文本数据，而 LDA 在长文本上的主题分析有很好的效果。另外，BTM 模型适用于短文本数据，其基于词对共现的主题分析对长文本来说是非常耗时的。

为了扩展短文本，首先本文对外部语料库做主题分析获得 LDA 主题模型 M_{LDA}；然后，根据模型 M_{LDA} 对每个数据块做主题推断，从而获取数据块中每个短文本的主题分布，有关主题推断的细节可以参考先前 Phan 等的工作[10]；最后，根据上步骤得到的主题分布，将主题下有代表性的词语添加到短文本中扩展短文本。为了降低噪音，本文选择前 N_t 个主题下 N_w 个代表性的词语扩展每一个短文本。

例如，给定一个短文本数据块 $\{d_j\}_{j=1}^{3}$，如图 3.2-(A)所示，图 3.2-(B)展示了根据主题模型 M_{LDA}，推断短文本数据块 $\{d_j\}_{j=1}^{3}$ 的主题分布结果 $\{\vartheta_j\}_{j=1}^{3}$，其中 ϑ_j 表示数据块中第 j 个短文本的主题分布，另外图 3.2-(A)提到的短文本数据来自 Snippets 数据集[10]。为了扩展短文本，本文假定一个主题的概率越高，其主题下代表性的词在短文本红出现的频次也就越高。因此，我们设定如果主题概率处于区间[0.5,1)，其代表词被扩展到短文本中三次，处于区间[0.2,0.5)和[0.1,0.2)，扩展频次分别为 3 和 2，处于区间[0.07,0.1)，其扩展频次为 1。另外，为了减少噪音的影响，我们选择不超过 N_t 主题（如 $N_t=3$），每个主题下选择 N_w 个代表词（如 $N_w=5$）去扩展短文本。图 3.2-(C)给出了用主题下的代表词扩展短文本数据块 $\{d_j\}_{j=1}^{3}$ 的结果，其扩展后的短文本数据块为 $\{d_j'\}_{j=1}^{3}$，其中相同主题下的词被省略了。

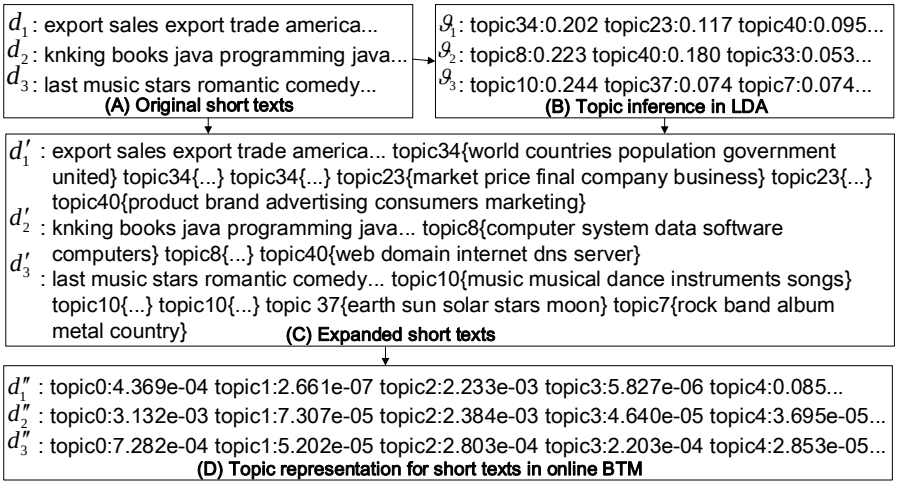

图 3.2 基于短文本扩展的在线 BTM 模型的例子

Fig 3.2 An example of online BTM based on short text expansion

3.2.2.3 短文本表示

根据上述方法完成短文本数据的扩展后，本章借助 OnlineBTM 将扩展后的短文本数据块中每个短文本表示为主题形式。OnlineBTM 是基于词对共现模式的针对短文本数据流的主题模型，在 t^{th} 时间下，其产生过程如图 3.3 所示：

1 Draw a topic distribution in the t^{th} time slice $\theta^{(t)} \sim Dirichlet(\alpha^{(t)})$;

2 For each topic k from K topics:

2.1 Draw a word distribution of topic in the t^{th} time slice $\phi_k^{(t)} \sim Dirichlet(\beta_k^{(t)})$;

3 For each biterm $b_j = \{w_{j,1}, w_{j,2}\} \in B^{(t)}$:

3.1 Draw a topic $z_j \sim Multinomial(\theta^{(t)})$;

3.2 Draw a word $w_{j,1}, w_{j,2} \sim Multinomial(\phi_{z_i}^{(t)})$.

图 3.3 在 t^{th} 时间下在线 BTM 的生成过程

Fig3.3 Generative process of online BTM in the t^{th} time slice.

其中 $B^{(t)} = \{b_j\}_{j=1}^{|B^{(t)}|}$ 表示在 t^{th} 时间片下的 biterm 集合，其大小用 $|B^{(t)}|$ 表示，$\alpha^{(t)}$ 和 $\beta_k^{(t)}$ 分别是针对 $\theta^{(t)}$ 和 $\phi_k^{(t)}$ 的 Dirichlet 参数，$\alpha^{(t)}=\{\alpha_k^{(t)}\}_{k=1}^{K}$ 是一个 K 维的多项式分布，$\beta_k^{(t)} = \{\beta_{k|w}^{(t)}\}_{w=1}^{W}$ 是 W 维的多项式分布，对称狄利克雷分布被用于初始化先验分布，即 $\alpha^{(1)} = (\alpha, \cdots, \alpha)$，$\beta^{(1)} = (\beta, \cdots, \beta)$。图 3.4 给出了 OnlineBTM 在 t^{th} 时间片下的图形表示。

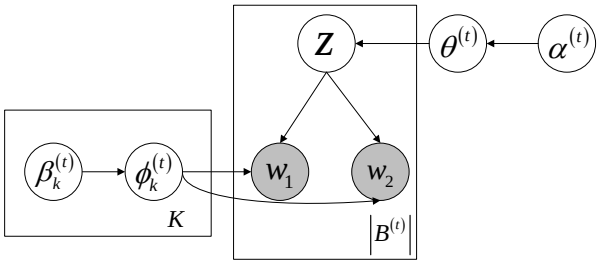

图 3.4 在时间片 t^{th} 下在线 BTM 的图形化表示

Fig 3.4 Graphic representation of online BTM in the t^{th} time slice.

我们借助 Gibbs 采样评估每个时间片下的 t^{th} 时间片下的主题分布和主题-词的多项式分布。t^{th} 时间片下数据块在经过 Gibbs 采样后，学得主题分布 $\theta^{(t)}=\{\theta_k^{(t)}\}_{k=1}^{K}$ 和主题-词的多项式分布 $\phi^{(t)}=\{\phi_k^{(t)}\}_{k=1}^{K}$，其中 $\phi_k^{(t)}=\{\phi_{k|w}^{(t)}\}_{w=1}^{W}$。假定短文本 d 包含 N_b 个无序词对，即 $\{b_j^{(d)}\}_{j=1}^{N_b}$，其中 $b_j^{(d)}=(w_{j,1}^{(d)},w_{j,2}^{(d)})$，则短文本 d 被分配到主题 k 的概率可以通过公式 3.1 表示：

$$P(k\mid d)=\sum_{j=1}^{N_b}P(k\mid b_j^{(d)})P(b_j^{(d)}\mid d) \tag{3.1}$$

其中 $P(k\mid b_j^{(d)})$ 表示短文本 d 中第 j 个词对 $b_j^{(d)}$ 被分配给主题 k 的概率，$P(b_j^{(d)}\mid d)$ 表示短文本 d 中包含词对 $b_j^{(d)}$ 的概率，可分别通过公式 3.2 和公式 3.3 计算获得：

$$P(k\mid b_j^{(d)})=\frac{\theta_k^{(t)}\phi_{k,w_{j,1}^{(d)}}^{(t)}\phi_{k,w_{j,2}^{(d)}}^{(t)}}{\sum_{k'}^{K}\theta_{k'}^{(t)}\phi_{k',w_{j,1}^{(d)}}^{(t)}\phi_{k',w_{j,2}^{(d)}}^{(t)}} \tag{3.2}$$

$$P(b_j^{(d)}\mid d)=\frac{n(b_j^{(d)})}{\sum_{i=1}^{N_b}b_i^{(d)}} \tag{3.3}$$

其中 $n(b_j^{(d)})$ 表示词对 $b_j^{(d)}$ 在短文本 d 中出现的频次。因此，通过上述过程可以获得短文本-主题的概率分布。另外，通过 Gibbs 采样，我们可以得到每个主题下词对的个数和每个词被分配给某个主题的次数，这里将主题 k 下词对的个数用符号 $n_k^{(t)}$ 表示，词 w 被分配给主题 k 的次数用符号 $n_{w|k}^{(t)}$ 表示，则下一个时间片下的超参数 $\alpha^{(t+1)}$ 和 $\{\beta_k^{(t+1)}\}_{k=1}^{K}$ 可以通过上述获得的 $n_k^{(t)}$ 和 $n_{w|k}^{(t)}$ 来调整，其中 $\alpha^{(t+1)}=\{\alpha_k^{(t+1)}\}_{k=1}^{K}$，$\beta_k^{(t+1)}=\{\beta_{k|w}^{(t+1)}\}_{w=1}^{W}$，计算方式如公式 3.4 和 3.5 所示

$$\alpha_k^{(t+1)}=\alpha_k^{(t)}+\lambda n_k^{(t)} \tag{3.4}$$

$$\beta_{k|w}^{(t+1)}=\beta_{k|w}^{(t)}+\lambda n_{w|k}^{(t)} \tag{3.5}$$

因此，针对 $(t+1)^{th}$ 时间片下的数据块，其产生过程与 t^{th} 时间片下的产生过程相似，采用新的超参数 $\alpha^{(t+1)}$ 和 $\{\beta_k^{(t+1)}\}_{k=1}^{K}$。

本章的方法借助 OnlineBTM 推断被扩展后的每个短文本的主题分布，因此，扩展后数据块 D' 中每个短文本就可以被表示为一组主题，即主题表示后的数据块 $D_i''=\{d_j''\}_{j=1}^{|D_i|}$，其中 $d_j''=\{z_{j,k}\}_{k=1}^{K}$。例如，图 3.2-(C)中被扩展的短文本 $\{d_j'\}_{j=1}^{3}$ 被表示成图 3.2-(D)中的一组主题 $\{d_j''\}_{j=1}^{3}$。

3.2.3 概念漂移检测

在短文本数据流中，主题随时间发生潜在的漂移称为概念漂移，其严重影响到分类的效果，因此如何适应短文本数据流中的概念漂移问题成为亟需解决的问题。为解决上述问题，本章设计一种基于主题分布的概念漂移检测方法。我们通过累加新的数据块中每个短文本与当前数据块的语义距离来判断新数据块相对于当前数据块是否发生概念漂移，如公式 3.6 所示：

$$dist(D''_{i+1}, D''_i) = 1/|D_{i+1}| \sum_{j=1}^{|D_{i+1}|} dist(d''_j, D''_i) \quad (3.6)$$

为计算公式 3.6 中的语义距离 $dist(d''_j, D''_i)$，根据类分布，首先将当前数据块划分成类簇 $\{I_c\}_{c=1}^{C}$，其中 $I_c = \{d''_l\}_{l=1}^{|I_c|}$，$|I_c|$ 表示 c^{th} 类的文本数目；然后计算短文本与所有类簇 $\{I_c\}_{c=1}^{C}$ 之间的语义距离；最后选择与某类簇语义距离最小的值表示短文本与当前数据块之间的语义距离，如公式 3.7 所示：

$$dist(d''_j, D''_i) = \min dist(d''_j, I_c) \quad (3.7)$$

其中 $dist(d''_j, I_c) = 1/|I_c| \sum_{l=1}^{|I_c|} dist(d''_j, d''_l)$， $d''_l \in I_c$

为了降低噪音的影响，如某个类簇可能会数量很少的短文本，在计算两个短文本的语义距离时，我们增加了权重，如公式 3.8 所示：

$$dist(d''_j, d''_l) = 1 - |I_c|/|D''_i| \cos(d''_j, d'') \quad (3.8)$$

其中 $\cos(d''_j, d''_l)$ 计算公式如 3.9 所示：

$$\cos(d''_j, d''z) = (z_{j,1} \cdot z_{l,1} + \cdots + z_{j,K} \cdot z_{l,K}) / (\sqrt{\sum_{k=1}^{K} (z_{j,k})^2} \cdot \sqrt{\sum_{k=1}^{K} (z_{l,k})^2}) \quad (3.9)$$

给定阈值 μ，判断是否发生概念漂移，如果 $dist(D''_{i+1}, D''_i) \in (\mu, 1]$，我们认为新数据块 D''_{i+1} 相对于当前数据块 D''_i 发生概念漂移。

3.2.4 集成模型的构建与更新

为了预测新的短文本，本章选择 H 个经过上述扩展并表示为主题的数据块分别构建基分类器。这里本章采用 SVM 作为基分类器，因为它广泛的被用于文本分类。当新的数据 D_e 块到来时，根据上述方法，我们首先借助外部语料库扩展获得 D'_e，然后将扩展后的数据块 D'_e 中每个短文本表示成一组主题 $D''_e = \{d''_j\}_{j=1}^{|D_e|}$。为了使用公式 3.1 预测新数据块 D''_e 中短文本 d''_j 的标签，我们需要获得该短文本 d''_j 相对于基分类器 f_h 的权值，该权值记为 $w_{h,j}$。根据公式 3.6 和 3.7，我们计算新数据块 D''_e 与构建基分类器 f_h 的数据块 D''_h 的语义距离 $dist(D''_e, D''_h)$，和短文本 d''_j 与数据块 D''_h 的语义距离 $dist(d''_j, D''_h)$，因此，权值 $w_{h,j}$ 可以通过公式 3.10 计算获得：

$$w_{h,j} = (1 - dist(d''_j, D''_h)) * (1 - dist(D''_e, D''_h)) \quad (3.10)$$

其中 $1 - dist(d''_j, D''_h)$ 表示数据块 D''_e 中短文本 d''_j 与数据块 D''_h 的语义相似性，$1 - dist(D''_e, D''_h)$ 表示两个数据块之间的语义相似性，用于减少概念漂移对分类结果

的影响。

为了更新集成模型 E，本章计算新的数据块 D_e'' 与用于构建集成模型的数据块之间的语义距离，同时新的数据块被用于构建新的分类器 f。当数据块 D_e'' 相对于集成模型中每个数据块均发生概念漂移，且集成模型 E 中基模型个数少于 H，则分类器 f 加入集成模型，如果等于，则用分类器 f 替换集成模型中最老的基模型。否则，分类器 f 被用于替换与其语义距离最小的数据块构建的基分类器。图 3.5 给出了整体流程。

Input:

String data D, the LDA topic model M_{LDA}, the ensemble model E, the set of H data chunks used for the ensemble model S, the threshold μ;

Output:

updated E, predicted labels in data chunk D_e;

1: **for** each data chunk D_e in D **do**
2: Expand D_e with the topic model M_{LDA} as D_e';
3: Represent expanded D_e' as D_e'' using online BTM;
4: **for** D_h'' **in** S **do**
5: **for** d_j'' **in** D_e'' **do**
6: Calculate semantic distance between d_j'' and D_h'' using Eq.(3.7);
7: **end for**
8: Calculate semantic distance between D_e'' and D_h'' using Eq.(3.6);
9: **end for**
10: **for** d_j'' **in** D_e'' **do**
11: Calculate weights using Eq.(3.10);
12: Predict d_j'' according to ensemble model E using Eq.(3.1);
13: **end for**
14: Detect concept drifts between D_e'' and each data chunk in S according to the threshold μ;
15: Build a new classifier f on D_e'' and update ensemble model E according to results of concept drifts;
16: **end for**

图 3.5 算法流程

Fig 3.5 Pseudocod of Our Approach

3.3 实验与分析

本节主要对比了本章所提算法与基准算法的实验结果，首先给出了实验中的短文本数据集和使用的基准算法，然后说明了实验的评价指标和参数设置，最后验证了本章所提算法在分类和概念漂移检测两方面的有效性。

3.3.1 数据集

我们选择了三个公开短文本数据集来验证本章所提算法的有效性，表 3.2 给出了数据集的相关信息。

表 3.2 数据集

Table 3.2 Data sets

数据集	类别	短文本数	数据集	类别	短文本数
Snippets	Business	1,500	News	Sport	8,189
	Computer	1,420		Business	5,365
	Culture-Arts-Ent	2,210		U.S	4,782
	Education-Science	2,660		Health	1,851
	Engineering	370		Sci_tech	2,870
	Health	1,180		World	6,254
	Politics-society	1,500		Entertainment	3,286
	Sports	1,420			
	总数	12,340			
Tweets	Arsenal	276,744			
	Blackfriday	34,481		总数	32,597
	Chelsea	340,194			
	Smartphone	152,194			
	总数	803,613			

（1）Snippets 数据集[3][10]：Snippets 是将预设定的表达放在网络搜索引擎上搜索的结果，总共 8 个类别，12340 条短文本。

（2）News 数据集：来源于 TagMyNews 数据集[4]，总共 7 类。每条新闻包含一个短标题，一个简短的描述以及一个日期等信息。实验中我们仅抽取短标题作为数据集，总共有 32597 条短文本。

（3）Tweets 数据集[46]：借助推特关键字追踪 API[5]，从推特中获取的 2012 年

[3] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz.

[4] http://acube.di.unipi.it/tmn-dataset.

[5] https://dev.twitter.com/docs/api/1.1/post/statuses/filter.

11 月到 12 月的推文数据，总共 4 类。实验中我们仅使用 2012 年 12 月的推文，按时间顺序获得 803613 条短文本数据。

上述三个基准数据集均采用 Stanford-Parser[6]进行分词处理。另外，为了验证本章所提概念漂移检测算法的有效性，我们重新组织 Snippets 数据集和 News 数据集，每个数据集有一个随机的概念漂移，其大小设置为 100 到 700 条短文本，另外我们为每个类别设置一个噪声因子 $r\%$，其值设为 5。

3.3.2 基准算法

本节给出了 4 个经典的短文本数据流分类算法和 5 个概念漂移检测算法用于验证本章所提算法在分类和概念漂移检测上的有效性，表 3.3 给出了基准算法的详细信息。上述所有的基准算法均来源于开源代码 MOA[47]。另外，本章所提基于文本扩展和概念漂移检测的短文本数据流分类方法用 OurE.Drift 表示，为了验证短文本扩展的有效性，我们将仅包含概念漂移检测的方法用 Our.Drift 表示。

表 3.3 基准算法

Table 3.3 benchmark algorithms

类别	方法	描述
数据流分类	Naive Bayes	一个以简单和低计算成本而闻名的分类器。
	Spegasos[48]	基于随机变量的Pegasos方法。
	KNN+PAW+ADWIN	概率近似窗口和自适应滑动窗口的k近邻。
	HoeffdingOptionTree[49]	基于流数据的决策选项树。
概念漂移检测	DDM[50]	概念漂移检测算法。
	CusumDM[51]	基于Cusum的漂移检测算法。
	PageHinkleyDM[52]	基于Page-Hinkley实验的漂移检测算法。
	HDDM_A_Test[53]	基于Hoeffding边界的在线漂移检测算法。
	HDDM_W_Test[53]	基于McDiarmid边界的在线漂移检测算法。

3.3.3 评价指标和参数设置

本章采用增量式的分类精度[27]来验证所提算法在短文本数据流上的有效性，假设数据流被划分为多个数据块，集成模型被训练预测新的数据块，同时用新数据块更新集成模型，因此我们可以获得在每个数据块上的分类精度，我们称之为增量式分类精度。为了验证本章所提概念漂移检测算法的有效性，我们使用衰落因子估计的先序错误率[52]作为评估方法，如公式 3.11 所示，

[6] http://nlp.standford.edu/software/standford-english-corenlp-2016-10-31-models.jar.

$$PreError(i)=\sum_{s=1}^{i}\tau^{i-s}e_s / \sum_{s=1}^{i}\tau^{i-s} \quad (3.11)$$

其中$\tau=0.995$表示衰落因子，$e_s=1$表示第s个短文本被错误分类，否则$e_s=0$。另外，我们引入三个统计评估方法，1）误报率：概念漂移检测中检测错误的概率；2）缺失率：存在概念漂移但未检测到的概率；3）延迟：概念漂移出现时，延迟检测到漂移的短文本数的平均值。

在参数设置上，两个主题模型的超参数α和β分别设置为0.5和0.01，其主题数统一设置为50，迭代次数为1000，OnlineBTM中的衰退值设置为0.5。在Snippets和News数据集上，数据块大小设为50，由于Tweets数据集太大，我们设置数据块大小为 1000。在概念漂移检测中，我们设置阈值$\mu=1-0.5/C$，其中C表示类别数。最后，我们选择来自libsvm[7]上的SVM作为基分类器，相应的参数为：选择C-SVC类型，线性核函数，其余均采用默认函数。另外，所有实验均执行在Inter Core i5处理器，频率为2.90GHz和8GB内存的一体机上。

3.3.4 实验结果与分析

本节验证了与基准算法对比，所提短文本数据流分类算法和概念漂移检测算法的有效性。

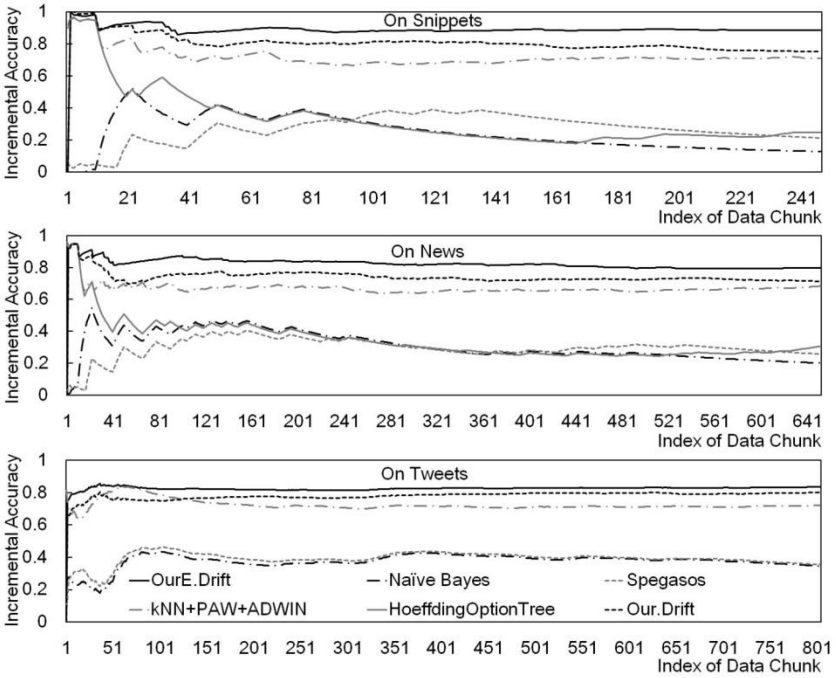

图 3.6 六种算法在3个数据集上的增量式精度

Fig 3.6 Incremental accuracy of 6 algorithms on 3 datasets.

[7] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

3.3.4.1 分类评估

本节对比了本章所提短文本数据流分类算法与基准算法在增量式分类精度上的效果，同时给出了时间性能上的对比。最后讨论了基模型数对所提算法的影响。

图 3.6 给出了本章所提算法与基准算法在三个数据集上的增量式分类精度（由于计算复杂度高，消耗内存过大，导致 HoeffdingOptionTree 算法不能在一周内处理，所以未给出），我们可以得出以下结论：1）与没有短文本扩展的 Our.Drift 算法相比，OurE.Drift 算法在 Snippets 和 News 数据集上平均高出 8%的增量式分类精度，在 Tweets 数据集上平均高出 4%的精度，具有明显的优势，说明了 OurE.Drift 算法上的短文本扩展是有效的。2）与 4 个传统数据流分类算法相比，OurE.Drift 算法在 3 个数据集上具有具有更高的分类精度，同时更加稳定。原因在于 4 个基准算法均构建一个增量式分类模型并没有考虑到短文本的稀疏性问题。而 OurE.Drift 算法借助基于短文本扩展的 OnlineBTM 缓解了短文本数据流的高维稀疏问题，增加了语义信息。另外，为适应短文数据流中的概念漂移问题，我们使用当前数据块与构建集成模型的数据块之间的语义距离作为对应基分类器的权值，而集成模型总是用最新的数据块更新。

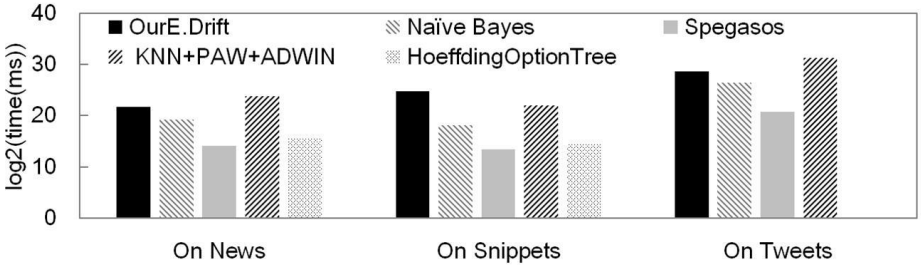

图 3.7 五种算法的耗时对比

Fig 3.7 Execution time of fivealgorithms.

图 3.7 对比了本章所提出的 OurE.Drift 算法与基准算法的时间性能。由实验结果可知，OurE.Drift 算法在 News 和 Tweets 数据集上处理速度快于 KNN+PAW+ADWIN 算法，但在 Snippets 数据集上更慢。原因分析如下：KNN+PAW+ADWIN 算法引入了概率近似窗口和自适应滑动窗口适应数据流中的概念漂移，当概念漂移出现时，需要频繁更新 K 的最近邻。本章所提算法的时间消耗主要在于 OnlineBTM 上，其时间复杂度为 $O(N_{iter}K|B^{(i)}|)$，$|B^{(i)}|=|M'|\bar{l}(\bar{l}-1)/2$，其中 N_{iter} 表示迭代次数，$|M'|$表示短文本数，$\bar{l}$ 表示短文本的长度。Snippets 数据集中短文本的平均长度超过 News 和 Tweets 数据集上的两倍，因此 OurE.Drift 算法在主题表示上花费了太多的时间。另外，与其他的基准算法相比，OurE.Drift 算法耗时更久。原因在于虽然我们降低短文本数据流的维度来降低集成模型构建和预测的时间，但在短文本扩展和主题表示上花费了太多的时间。

图 3.8 给出了基模型数从 2 个到 15 个对分类性能影响的实验结果。由图可知，随基模型数的增加，Snippets 和 Tweets 数据集上的 OurE.Drift 算法的增量式分类精度先呈现出增加然后在基模型数大于或等于 10 的时候趋于平稳，而 News 数据集上的实验结果却出现下降的趋势。原因在于基模型数越大，预测结果的多样性越大，而 News 数据集上的特征空间更加稀疏，不同类之间的特征重合更明显，因此随基模型数的增加，其增量式分类精度呈现出下降的趋势。

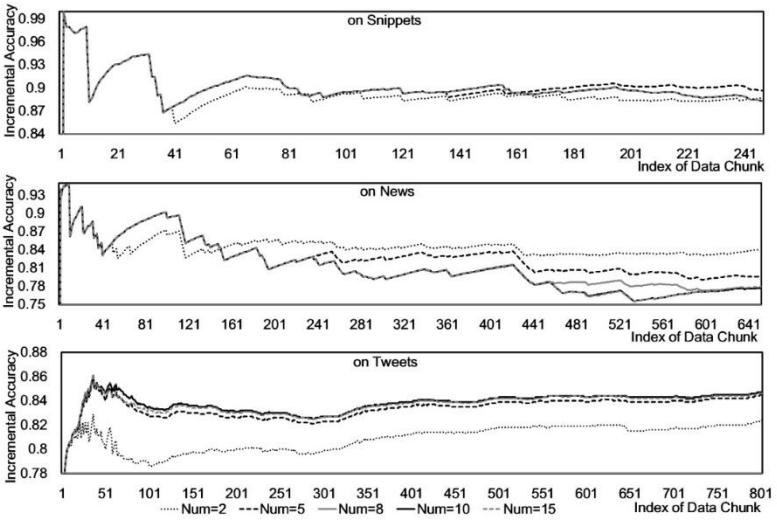

图 3.8 基模型数在 3 个数据集上的的影响

Fig 3.8 Effects of the number of base classifiers on 3 datasets.

3.3.4.2 概念漂移检测评估

本节首先给出了 OurE.Drift 算法与 KNN+PAW+ADWIN 算法的先序错误率，其中 KNN+PAW+ADWIN 算法是基准数据流分类算法中表现突出的。然后，展示了本章所提概念漂移检测算法与基准漂移检测算法的统计数据。

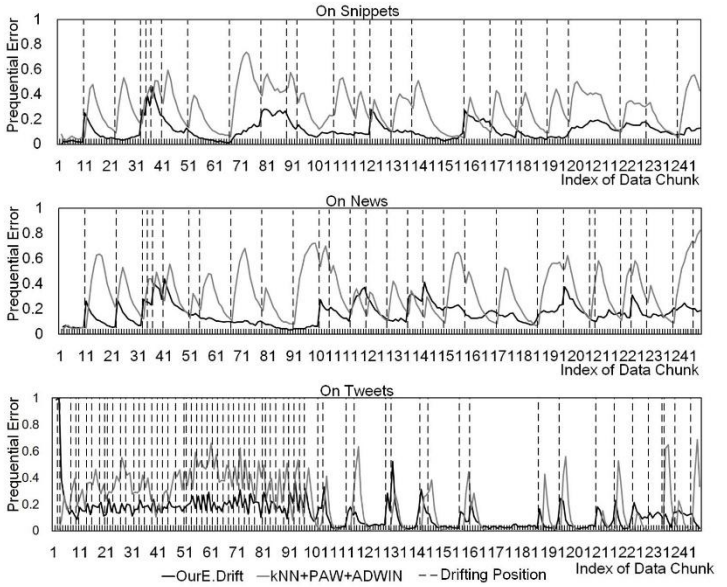

图 3.9 OurE.Drift 与 KNN+PAW+ADWIN 的先序错误率

Fig 3.9 Prequential Error between OurE.Drift and KNN+PAW+ADWIN.

图 3.9 提供了 OurE.Drift 算法与 KNN+PAW+ADWIN 算法的先序错误率评估结果。由此可观察到以下情况，1）当数据流出现概念漂移时，OurE.Drift 算法比 KNN+PAW+ADWIN 算法有更低的先序错误率，原因在于 OurE.Drift 算法使用多个不同的数据块分别构建基分类器，包含更多的已经发生的概念，更能有效的适应漂移现象。2）OurE.Drift 算法比 KNN+PAW+ADWIN 算法能更早的从概念漂移中恢复，主要原因是如果检测到概念漂移，由最新的数据块构架的基分类器会快速的更新集成模型适应漂移。

表 3.4 漂移检测评估

Table 3.4 Drifting detection statistics (FA: False Alarm, Miss: Missing)

Measures	**DDM**	**Cusum-DM**	**PageHinkleyDM**	**HDDM_A_Test**	**HDDM_W-Test**	**ours**
Snippets						
FA(%)	5.26(3)	6.25(4)	**0(1)**	7.69(5)	4.17(2)	**0(1)**
Miss(%)	28(4)	37.5(5)	96.0(6)	**4.0(1)**	8.0(2)	12(3)
Delay	32.89(4)	64.47(5)	122.0(6)	9.625(2)	11.43(3)	**0(1)**
News						
FA(%)	**0(1)**	**0(1)**	**0(1)**	1.79(2)	1.79(2)	8.75(3)
Miss(%)	64.18(4)	62.69(3)	94.03(5)	17.91(2)	17.91(2)	**6.49(1)**
Delay	37.5(4)	82.32(5)	243.0(6)	7.93(2)	12(3)	**0(1)**
Tweets						
FA(%)	3.79(3)	1.98(2)	**1.12(1)**	5.20(5)	7.88(6)	5.13(4)
Miss(%)	56.60(3)	62.57(5)	96.54(6)	57.96(4)	52.15(2)	**16.94(1)**
Delay	133.72(3)	131.23(2)	468.92(6)	166.59(4)	158.82(5)	**1.92(1)**

表 3.4 展示了本章所提漂移检测算法与基准检测算法的统计信息评估结果。基准检测算法的结果均选择于上述所提的数据流分类算法中最好的结果。表中括号里的数字表示所有检测算法的结果排序，加粗数字表示对应检测算法效果最好。由表中统计数据我们可以得到以下结论：1）针对 News 和 Tweets 数据集，与基准检测算法相比，本章所提检测算法在缺失率和延迟评估上效果最好。主要因为我们借助外部语料库扩展短文本缓解短文本数据流的稀疏性问题，同时基于主题分布检测概念漂移，因此所提检测算法对概念漂移问题更加敏感。HDDM_A/W_Test 在 Snippets 数据集上的缺失率最低，这是因为非加权/加权的统计方法被应用到 Hoeffding 边界估计中，它要求短文本又更多的统计信息，而 Snippets 数据集语义信息充足。2）PageHinkleyDM 总是保持最低的错误率，由于其建立在至少 90% 的

漂移未被检测到的前提下。原因在于短文本数据流的特征高维稀疏问题使其很难区分观测值与其均值之间的差异。

3.4 本章小结

本章针对 OnlineBTM 存在的问题，提出一种基于短文本文本扩展和概念漂移检测的短文本数据流分类算法。首先，我们借助来自 Wikipedia 的外部语料库扩展短文本用以缓解短文本数据流的稀疏性问题，利用 OnlineBTM 将扩展后的短文本数据表示成主题从而解决短文本数据流的高维问题。然后为了检测短文本数据流中的概念漂移问题，本章提出基于主题分布的漂移检测算法，同时使用多个基分类器建立集成模型，借助概念漂移结果更新模型。最后大量的实验结果表明，与传统的方法相比，本章所提方法可以在短文本流分类和概念漂移检测中获得更好的性能。

然而，本章所提方法属于有监督算法，需要大量的标记短文本数据，而标记数据的获取耗费大量的人力和财力，且获取数量十分有限。下一章将进一步研究解决类标签缺失问题的半监督短文本分类方法。

第四章　基于标签传播的半监督短文本数据流分类算法研究

实际应用中的类标签缺失问题导致有监督短文本数据流分类算法难以被广泛应用。虽然有标记的短文本数据稀少，但海量的无标记数据同样包含着丰富的语义信息，因此，受到半监督思想的启发，本章提出一种基于标签传播的半监督短文本数据流分类算法，通过借助少量有标记的短文本数据和丰富的无标记短文本数据共同训练集成模型从而提高分类精度。

4.1 引言

现在流行的社交网络平台如：Twitter、微博、博客等，每天都会产生海量的短文本数据流，其中蕴含着重要的商业价值与研究价值。因而如何获取短文本数据流中潜在的、有价值的信息对于企业、政府决策以及个人学习变得尤为重要。短文本分类问题作为基础研究倍受研究者的关注。目前针对短文本分类的研究主要集中于有监督学习，然而实际应用中有标记短文本数据的获取不仅消耗大量的人力、物力、财力，且获取的数量有限，而无标记短文本数据的获取相对简单，且一般数量巨大，因此，基于半监督的短文本分类越来越受到研究者的关注。

为了提升分类效果，基于半监督的分类方法试图借助大量易获取的无标记数据和少量昂贵的标记数据来共同训练分类器。目前，已有的基于半监督学习的短文本分类方法主要包括：Yin 等[30]提出基于半监督学习和 SVM 的短文本分类方法；Li 等[32]提出了一种基于融合相似性测量和类中心的半监督短文本分类方法。上述两种方法都是通过迭代的方式为无标记数据打上类标签来提升分类性能，然而如果某个短文本被打上错误的标签信息就有可能影响到其他未标记的短文本。De Silva 等[31]利用无标记数据构建相似矩阵，同时与分类器结合提出一种半监督学习框架；Widmann 等[33]为缓解短文本的稀疏性问题，提出一种基于图表示的半监督短文本分类方法。以上两种方法虽然借助无标记数据提升了短文本的分类性能，然而并未解决短文本数据流中存在的特征高维稀疏问题。此外，实际短文本数据流除具有上述所提的特征高维稀疏问题外，还具有海量、快速、概念漂移等特点，这将导致上述所提的批处理算法难以适应。

为此，研究者探索了基于半监督学习的数据流分类方法，其主要工作有：Zhang 等[34]提出一种基于分类器与聚类模型的集成分类方法，然而该集成方法仅通过其本身的集成特性去适应概念漂移，未考虑概念漂移检测，因此，当概念漂移发生时，该方法很难快速的适应，导致分类性能降低。Li 等[35]为解决不完全标记数据流和重复概念漂移现象，提出一种基于增量式单决策树模型的半监督分类算法。Zhu 等[37]提出基于线最大流算法，提出一种新的基于半监督数据流分类方法。然

而上述方法都是针对特征低维稠密的数据流分类问题，难以适应特征高维稀疏的短文本数据流分类问题。

因此，本章提出了一种基于标签传播的半监督短文本数据流分类算法。首先，为了降低短文本数据流存在的特征高维稀疏问题，该算法首先从 Wikipedia 中获取外部语料，并借助词向量训练工具（Word2Vec[8]）训练相应语料获得原始词向量集合用以丰富短文本的语义信息。其次，在有标记和无标记的短文本数据上分别训练分类器和聚类器构建集成模型，并设计一种基于簇相似度的计算方法将已知标签信息传播到聚类簇中，从而使得聚类簇也可以预测标签。此外，为适应概念漂移，提出一种基于聚类簇的概念漂移检测机制。最后，在三组基准短文本数据流上的实验表明：所提算法取得了良好的分类精度和宏平均，同时能快速适应短文本数据流中的概念漂移。

4.2 算法描述

本节详细介绍了基于向量表示和标签传播的半监督短文本数据流分类算法，首先给出问题定义和算法的基本框架；进而详细介绍算法框架的技术细节，包括基于 Word2Vec 的特征向量表示、有标签数据到聚类簇的标签传播、概念漂移检测以及集成模型的构建与更新。

4.2.1 问题定义

按有无标签将给定的短文本数据流 S 中的文本划分为数据块集合，记为 $S=\{D_1^l, D_2^l, \cdots, D_m^l, D_{m+1}^u D_{m+2}^u; \cdots, D_{m+n}^u\}$，其中 l 和 u 分别表示标签已知和未知，m，n 分别表示有标记短文本数据块和未标记短文本数据块的个数。$D_i^x=\{d_{i,1}^x, \cdots, d_{i,m'}^x, \cdots, d_{i,M}^x\}$，$x\in\{l,u\}$，其中 D_i^x 表示第 i 个已标记或未标记的数据块，共包含 M 个短文本。$d_{i,m'}^x=(W_{i,m'}^x, y_{i,m'}^x)$ 表示第 i 个标记或未标记数据块中第 m' 个短文本，其中 $W_{i,m'}^x$ 表示第 i 个数据块的第 m' 个短文本的词集，记为 $W_{i,m'}^x=\{w_{i,1}, \cdots, w_{i,|d_{i,m'}^x|}\}$，$|d_{i,m'}^x|$ 表示 $W_{i,m'}^x$ 的单词数，$y_{i,m'}^x\in Y=\{c_1, \cdots, c_r\}$ 表示第 i 个数据块的第 m' 个短文本的类别标签，r 表示标签的个数。

图 4.1 给出了本章所提算法的基本框架图，首先，由于第三章所提基于文本扩展的 OnlineBTM 模型的主题表示不适用于标签传播，本章设计一种新的基于 Word2Vec 的词向量表示方法，即借助 Word2Vec 训练外部语料库获得丰富的原始词向量集合，然后当短文本数据块到来时，利用已获得的原始词向量集合将数据块中的每个短文本进行向量表示用以扩展短文本的语义信息，缓解稀疏性问题，同时降低特征维度，我们将表示后的短文本数据流记为

[8] https://radimrehurek.com/gensim/models/word2vec.html.

$S'=\{D_1'^l,D_2'^l,\cdots,D_m'^l,D_{m+1}'^uD_{m+2}'^u,\cdots,D_{m+n}'^u\}$，其中$D_i'^x=\{d_{i,1}'^x,\cdots,d_{i,m}'^x\}$，$d_{i,m'}'^x=\{V_{i,m'}^x,y_{i,m'}^x\}$，$V_{i,m'}^x=\{v_1,\cdots,v_F\}$表示一组$F$维特征向量。为适应短文本数据流中类标签的大量缺失问题，我们从数据流S'中选择最新的a个标记数据块构建a个分类器和$k-a$个未标记数据块构建聚类器，其中k个数据块记为$\{D_1'^l,\cdots,D_a'^l,D_{a+1}'^u\cdots,D_k'^u\}$，对应的$k$个基模型记为$E=\{\lambda^1,\cdots,\lambda^a,\lambda^{a+1},\cdots,\lambda^k\}$。另外本章将$k-a$个未标记数据块构建的聚类模型对应的聚类簇集合记为$\{G_{a+1}^u,\cdots,G_j^u,\cdots,G_k^u\}$，其中$G_j^u=\{g_{j,1}^u,\cdots,g_{j,h}^u,\cdots,g_{j,r}^u\}$，$g_{j,h}^u$表示第$j$聚类模型中第$h$个簇，本章设置聚类簇个数与短文本数据的类标签个数相同，均用r表示。

当新的短文本数据块$D_{k+1}^x=\left\{d_{k+1,1}^x,\cdots,d_{k+1,m'}^x,\cdots,d_{k+1,M}^x\right\}$到来时，首先，借助原始词向量集合将$D_{k+1}^x$中的每个短文本表示为特征向量$D_{k+1}'^x=\left\{d_{k+1,1}'^x,\cdots,d_{k+1,m'}'^x,\cdots,d_{k+1,M}'^x\right\}$。基于聚类簇的概念漂移检测算法计算数据块$D_{k+1}^x$相对于$k$个短文本数据块$\{D_1'^l,\cdots,D_a'^l,D_{a+1}'^u\cdots,D_k'^u\}$是否发生概念漂移，根据概念漂移结果构建集成模型$E$。最后基于集成模型$E$，利用公式 4.1 预测新数据块$D_{k+1}'^x$中每个短文本$d_{k+1,m'}'^x$的标签信息：

$$\begin{aligned} y^* &= argmax_{y\in Y}P(y\mid d_{k+1,m'}'^x,E) \\ &= argmax_{y\in Y}\sum_{i=1}^{k}w_iP(y\mid d_{k+1,m'}'^x,\lambda_i) \end{aligned} \tag{4.1}$$

集成模型E包含标记数据块训练的a个分类器和未标记数据块训练的$k-a$个聚类器。其中a个分类器可以直接预测新数据块中每个短文本的类标签，但$k-a$个聚类模型预测短文本得到的是聚类簇 ID 而非真正的类标签，因此本章设计出一种基于簇相似度的计算方法，将标记短文本数据的簇信息传播到聚类模型中每个聚类簇中，即基于簇的标签传播。通过上述的标签传播方法使得聚类簇能够获取标签信息从而预测短文本。因此，$P(y\mid d_{k+1,m'}'^x,E)$就可以改写为如下形式：

$$P(y\mid d_{k+1,m'}'^x,E)=\sum_{i=1}^{a}w_iP(y\mid d_{k+1,m'}'^x,\lambda_i)+\sum_{j=a+1}^{k}\sum_{h=1}^{r}w_jP(y\mid g_{j,h}^u)P(g_{j,h}^u\mid d_{k+1,m'}'^x,\lambda_j) \tag{4.2}$$

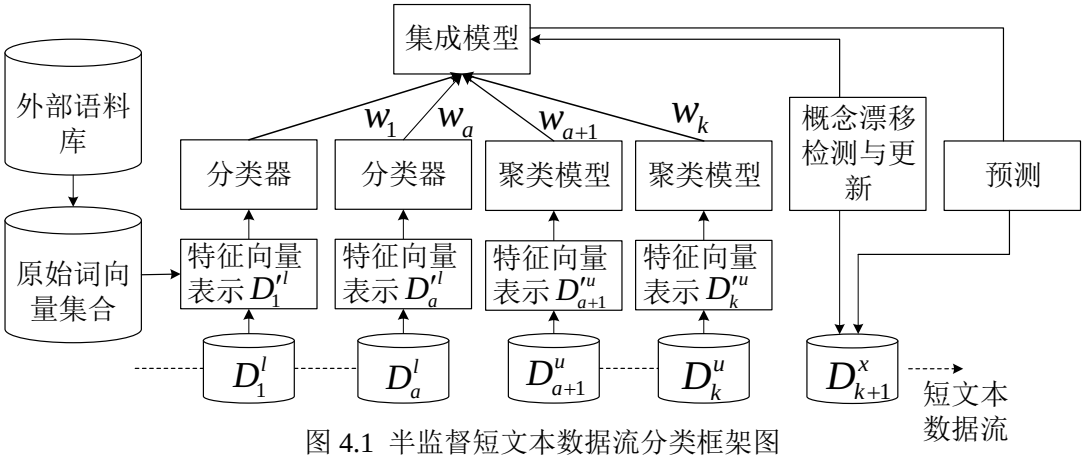

图 4.1 半监督短文本数据流分类框架图

Fig 4.1 Framework of semi-supervised short text streams classification.

4.2.2 基于 Word2Vec 的词向量表示

短文本数据流的特征高维稀疏问题导致传统的词袋模型（即在词袋模型中文本被视为无序的词汇集合，其语法与词汇之间的关系被忽略。）很难适应。第三章所提算法设计了一种基于短文本扩展的 OnlineBTM 模型，该方法虽然借助来自 Wikipedia 的外部语料库扩展短文本，然后借助 OnlineBTM 主题模型将扩展后的短文本表示成主题从而缓解了短文本数据流的特征高维稀疏问题，但由于扩展的词数有限导致扩展后的短文本仍属于短文本，其语义信息不充足，如果直接应用到本章算法中会导致后续的标签传播算法效果不理想，如图 4.6 所示。因此本章所提算法采用 Word2Vec 工具训练外部语料库获得语义丰富的原始词向量集合，然后借助训练好的原始词向量集合将短文本数据流中每个短文本映射成一组特征向量，从而丰富每条短文本的语义信息。

本章使用 2017 年 12 月 20 日至 2018 年 1 月 1 日更新发布的全部维基百科数据集[9]作为外部语料库。词向量训练工具 Word2Vec 被用于训练外部语料库获得原始词的向量集合，这里将训练完成的原始词向量集合记为 $C=\{(w'_1,V'_1),\cdots,(w'_{n'},V'_{n'}),\cdots,(w'_{|C|},V'_{|C|})\}$，其中 $V'_{n'}=\{v'_{n',1},\cdots,v'_{n',F}\}$ 表示第 n' 个词 $w'_{n'}$ 的向量集合，向量维度为 F，$|C|$ 表示外部语料库中总的单词数。为将短文本数据流中的每条短文本进行特征向量表示，我们首先根据原始词向量集合 C 找到短文本中每个词对应的词向量，然后然后根据每个词的词向量和每个词在短文本中出现的概率计算短文本的特征向量。例如，给定数据块 D_i^x 中的短文本 $d_{i,m'}^x$，其词集为 $W_{i,m'}^x=\{w_{i,1},\cdots,w_{i,|d_{i,m'}^x|}\}$，从原始词向量集合 C 中找到词集 $W_{i,m'}^x$ 中每个词对应的词向量记为 $\{V_{i,1},\cdots,V_{i,|d_{i,m'}^x|}\}$，则短文本 $d_{i,m'}^x$ 的特征向量计算入公式 4.3 所示：

$$V_{i,m'}^x=\sum_{b=1}^{|d_{i,m'}^x|}P(w_{i,b}\,|\,W_{i,m'}^x)\times V_{i,b} \tag{4.3}$$

其中 $P(w_{i,b}\,|\,W_{i,m'}^x)$ 表示词 $w_{i,b}$ 在词集 $W_{i,m'}^x$ 中出现的概率。因此，短文本 $d_{i,m'}^x$ 可以被表示为向量，即向量表示后的短文本为 $d'^x_{i,m'}=\{V_{i,m'}^x,y_{i,m'}^x\}$。通过上述特征表示方法，我们可以将数据块 D_i^x 中所有短文本进行特征向量表示，从而获得数据块 $D'^x_i=\{d'^x_{i,1},\cdots,d'^x_{i,m'},\cdots,d'^x_{i,M}\}$。

本节通过外部语料库训练的原始词向量集合中的词向量表示短文本数据流中对应的词，扩展了短文本的语义信息，缓解短文本数据的稀疏问题，同时将短文本表示为给定的 F 维特征向量降低了短文本的特征维度，解决了短文本数据流的稀疏性问题。

[9] https://dumps.wikimedia.org/enwiki/20180101/

4.2.3 基于簇的标签传播

聚类模型预测短文本获得的是聚类簇 ID 而非真正的类标签，因此，必须为聚类模型中的每个聚类簇寻找合适的类标签才能预测新的短文本。本章所提算法通过基于簇的相似度计算方式，利用构建分类器的 a 个标记数据块的簇信息预测 $k-a$ 个聚类模型中聚类簇的类标签。具体流程为：首先给定 a 个经过 4.2.2 节特征向量表示的有标记数据块 $\{D_1'^l,\cdots,D_i'^l,\cdots,D_a'^l\}$，根据类标签信息，将每个标记数据块划分为 r 个簇，例如第 i 个标记数据块 $D_i'^l$ 的簇信息记为 $G_i^l=\{g_{i,1}^l,\cdots g_{i,q}^l,\cdots g_{i,r}^l\}$。然后通过公式 4.4 将标记数据块的簇信息传播给聚类模型的聚类簇，从而每个聚类簇就有了类标签，即聚类模型可以预测新的短文本数据的类标签。其中，标记数据块 $D_i'^l$ 中簇 $g_{i,q}^l$ 属于类标签 y 的概率 $P(y|g_{i,q}^l)$ 为先验概率，而 $P(g_{i,q}^l|g_{j,h}^u)$ 概率本节通过一种基于簇相似度的方法计算两个簇之间的相似度，即 $P(g_{i,q}^l|g_{j,h}^u)\propto sim(g_{i,q}^l,g_{j,h}^u)$。

$$P(y|g_{j,h}^u)=\sum_{i=1}^{a}\sum_{q=1}^{r}P(y|g_{i,q}^l)P(g_{i,q}^l|g_{j,h}^u) \tag{4.4}$$

本节通过计算两个簇的半径之和与两个簇的中心距离的比值表示两个簇之间的相似性，形式化表示为公式 4.5：

$$sim(g_{i,q}^l,g_{j,h}^u)=(r_{iq}+r_{jh})/(\|\bar{u}_{iq},\bar{u}_{jh}\|_2^2) \tag{4.5}$$

其中，$\bar{u}_{iq}$ 和 $\bar{u}_{jh}$ 分别表示簇 $g_{i,q}^l$ 和 $g_{j,h}^u$ 的中心点，可定义为：$\bar{u}_{iq}=1/|g_{i,q}^l|\sum_{m'=1}^{|g_{i,q}^l|}V_{i,m'}^l$，$r_{iq}$ 和 r_{jh} 分别表示两簇 $g_{i,q}^l$ 和 $g_{j,h}^u$ 对应的半径，可定义为 $r_{iq}=1/|g_{i,q}^l|\sum_{m'=1}^{|g_{i,q}^l|}\|V_{i,m'}^l,\bar{u}_{iq}\|_2^2$，其中 $\|\|_2^2$ 表示平均欧氏距离，$|g_{i,q}^l|$ 表示构成 $g_{i,q}^l$ 的短文本数量。因此，$sim(g_{i,q}^l,g_{j,h}^u)$ 的值取决于两个簇的半径之和与两个簇的中心距离的可能关系。当 $0<sim(g_{i,q}^l,g_{j,h}^u)\le 1$，表示两个簇的半径之和小于两个簇的中心距离，则簇 $g_{i,q}^l$ 和 $g_{j,h}^u$ 不相似；当 $sim(g_{i,q}^l,g_{j,h}^u)>1$，表示两个簇的半径之和大于两个簇的中心距离，则簇 $g_{i,q}^l$ 和 $g_{j,h}^u$ 相似，且 $sim(g_{i,q}^l,g_{j,h}^u)$ 越大，相似性越高。

4.2.4 基于聚类簇的概念漂移检测

短文本数据流往往存在隐藏的概念漂移，严重影响分类的效果，因此如何检测概念漂移成为亟待解决的重要问题。本节设计了一种基于聚类簇的概念漂移检测方法，其具体流程如下：

（1）将经过特征向量表示后的新数据块 $D_{k+1}'^x$ 聚类，获得聚类簇 $G_{k+1}'^x=\{g_{k+1,1}^x,\cdots,g_{k+1,p}^x,\cdots,g_{k+1}^x\}$；

（2）计算数据块 $D_{k+1}'^x$ 中的聚类簇 $g_{k+1,p}^x$ 与用于构建集成模型 E 中数据块 $D_i'^x$ 中的每个簇 $G_i^x=\{g_{i,1}^x,\cdots,g_{i,q}^x,\cdots,g_{i,r}^x\}$ 的语义相似性，如公式 4.6 所示，为了减少数据

块 $D_i'^x$ 中可能存在某个簇有很少的短文本的影响，本节在 $sim(g_{k+1,p}^x, D_i'^x)$ 计算时设置权值；

$$sim(g_{k+1,p}^x, D_i'^x) = max_{g_{i,q}^x \in G_i^x}(|g_{i,q}^x| / \sum_{h=1}^r |g_{i,h}^x| \times sim(g_{k+1,p}^x, g_{i,q}^x)) \quad (4.6)$$

（3）通过公式 4.7 计算新到来的数据块 $D_{k+1}'^x$ 与当前用于构建集成模型 E 中数据块 $D_i'^x$ 的语义距离，

$$dist\left(D_{k+1}'^x, D_i'^x\right) = 1/\sum_{h=1}^r |g_{k+1,h}^x| \times \sum_{p=1}^r (|g_{k+1,p}^x| / sim(g_{k+1,p}^x, D_i'^x)) \quad (4.7)$$

为了检测新的数据块 $D_{k+1}'^x$ 相对于用于构建集成模型的数据块是否发生概念漂移，与前一张章相同，设置概念漂移检测阈值 μ，如果 $dist(D_{k+1}'^x, D_i'^x) \in (\mu, 1]$，则表示新的数据块 $D_{k+1}'^x$ 相对于数据块 $D_i'^x$ 发生概念漂移。

4.2.4 集成模型的构建与预测

为设置集成模型中每个基模型的权值，本节通过计算新的数据块 $D_{k+1}'^x$ 的聚类簇与当前用于构建集成模型 E 的数据块的聚类簇的相似性来获得。例如公式 4.8 计算的是基模型 λ_i 的权值 w_i。另外如果基模型 λ_i 是分类器，则 Z 为分类器的权值之和，即为 $Z = \sum_{i=1}^a w_i$，否则为聚类器的权值之和，即为 $Z = \sum_{i=a+1}^k w_i$.

$$w_i = \frac{1}{Z}\begin{Bmatrix} sim\left(g_{i,1}^x, g_{k+1,1}^x\right) & \cdots & sim\left(g_{i,1}^x, g_{k+1,r}^x\right) \\ \vdots & \vdots & \vdots \\ sim\left(g_{i,r}^x, g_{k+1,1}^x\right) & \cdots & sim\left(g_{i,r}^x, g_{k+1,r}^x\right) \end{Bmatrix} \quad (4.8)$$

为了预测新的短文本数据块 D_{k+1}^x，我们首先训练最新的 a 个有标记数据块获得分类器，同时选择最新的 $k-a$ 个无标记数据块构建的聚类模型，这里的分类器采用 SVM（SVM 分类器是适合短文本数据的），聚类方法使用 Kmeans；然后借助由外部语料库获取的原始词向量集合将新的短文本数据块 D_{k+1}^x 进行特征向量表示获得 $D_{k+1}'^x$；接着检测新数据块 $D_{k+1}'^x$ 相对构建集成模型 E 的 k 个数据块是否发生概念漂移，如果新的数据块 $D_{k+1}'^x$ 相对于 k 个数据块中某个数据块发生概念漂移，则该数据块构建的基模型不参与集成模型 E 的构建；最后，利用公式 4.1 预测新数据块 $D_{k+1}'^x$ 中每个短文本的类标签。

4.3 实验与分析

本节主要对比了本章所提算法与基准算法的实验结果，首先给出了实验中的短文本数据集和评价指标，然后说明了实验的基准算法和参数设置，最后验证了本章所提算法在半监督分类和概念漂移检测两方面的有效性。

4.3.1 实验数据集和评价指标

本章所提算法使用到的实验数据集和第三章的数据集相同，Snippets 数据集通过随机打乱调整获得，News 和 Tweets 数据集均根据时间排序获得，另外本节同样

采用 Stanford-Parser 对三个基准数据集进行分词处理。为了模拟半监督短文本数据流分类，本节为每个数据集设置标签缺失率（即无标记短文本占整个数据集的比率。），表 4.1 总结了三组数据集的数据规模、特征维、标签个数与标记缺失率，前两个数据集划分的数据块大小为 500，而推文数据集相对规模较大，划分的数据块大小为 1000。另外，人工设置的标记缺失率为 80%。

表 4.1 实验使用的数据集

Tabel 4.1 Data set used in the experiment

数据集	尺寸	特征维	标签数	标记缺失率
搜索片段	12000	22000	8	80%
新闻标题	32500	23600	7	80%
推文	803000	290000	4	80%

为了评估本章所提算法在分类上的有效性，我们使用的评价指标如下：1）分类精度：指预测正确的短文本数与预测短文本总数的比值。2）宏平均（MacroF1）：评价所有类别的综合指标，计算方式如公式 4.9 所示，准确率（Precision）表示被正确分为某类的短文本数与分为该类短文本总数的比值，召回率（Recall）表示正确分为某类的短文本数与预测数据块中真实属于该类的短文本总数的比值。另外，为评估所提概念漂移检测算法的有效性，本节使用上章所用的衰落因子估计的先序错误率[52]。

4.3.2 基准算法和参数设置

为验证本章所提基于标签传播的半监督短文本数据流分类算法的有效性，本节选择了四个基准算法，包括两个半监督数据流分类算法：基于分类器与聚类模型的集成分类方法 ECU[34]与用于解决重复概念漂移问题的基于增量式单决策树模型的半监督数据流分类算法 REDLLA[35]，以及两个批处理算法 TriTrain[54]和 SVM[3]。其中 TriTrain 是由 Zhou 等提出的一种基于协同训练的半监督分类算法。实验选取了前 a 个标记短文本数据块和 $k-a$ 个未标记短文本数据块构建模型来预测短文本数据。SVM 属于有监督分类算法，实验选取前 a 个标记短文本数据块构建一个 SVM 模型用于预测短文本数据。关于基准算法所选择的分类器与聚类模型问题：由于 REDLLA 算法是基于单决策树构建分类器，同时对叶子节点进行聚类，所以它采用 J48 决策树作为分类器。为与本章所提算法做对比，ECU 和 TriTrain 均采用 SVM 作为分类器，最后基准算法所涉及的聚类器均使用 Kmeans。另外四个基准算法均使用 TF-IDF 表示的短文本数据作为输入，为验证所提短文本的特征向量表示有利于所提算法，实验将所提算法的输入由特征向量表示的短文本改为 TF-IDF 表示的短文本进行实验，本章将其称为 BOG，所提算法称为 SSC。

在参数设置上，基模型个数$k=10$，特征向量维度$F=100$，其中分类器个数$a=2$。另外，检测概念漂移的阈值$\mu=1-0.5/r$。上述 SVM 分类器均来自于 libsvm，选择的 SVM 类型为 C-SVM，其惩罚系数为 32，使用线性核函数，其余均为默认值。测试环境与上章相同，是基于 Intel Core i5 处理器，2.90 GHz 频率和 8G 内存的一体机。

4.3.3 实验结果与分析

本节主要考察本章所提基于标签传播的半监督短文本数据流分类算法的两大实验性能，一是与实验基准算法相比，本章所提算法 SSC 在短文本数据流上的分类性能与时间开销。二是考察 SSC 算法中基于簇的概念漂移检测机制的有效性。另外，由表 4.1 可以看出推文数据集的原始特征维度过高，如果采用 Kmeans 算法聚类会消耗过高时间，因此 ECU、REDLLA 以及 BOG 算法仅在前两个数据集上运行。最后，给出了第三章所提基于短文本扩展的 OnlineBTM 主题模型的表示方法与本章所提基于 Word2Vec 的特征表示方法的对比实验结果。

4.3.3.1 分类性能评估

本节首先介绍了与基准算法对比，本章所提算法在 3 个短文本数据集上的分类结果。

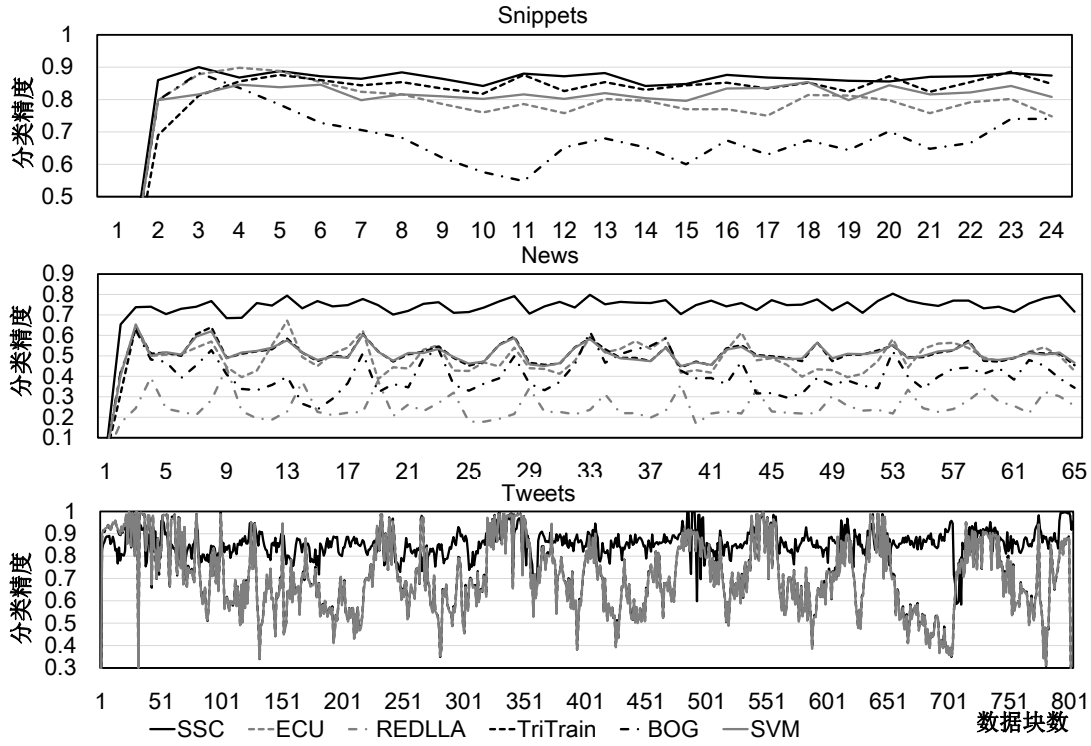

图 4.2 SSC 与基准算法在分类精度上的对比

Fig 4.2 Experimental comparison between SSC and baseline algorithms on accuracy.

图 4.2 和图 4.3 展示了 SSC 与基准算法在三个数据集上的分类的精度和宏平均

的实验结果，其中第一个数据块未被预测。由实验结果可得如下结论：

（1）本章所提算法 SSC 在分类精度上显著优于 BOG 算法，这是因为 SSC 算法中采用的 Kmeans 聚类模型比较适于低维稠密特征空间的聚类，为了分类短文本数据流，SSC 算法借助外部语料库的词向量信息丰富了短文本的语义信息，从而缓解了特征高维稀疏性。而基于词袋模型的 BOG 算法是通过利用原始词的频率表示短文本，未考虑短文本数据流存在的特征高维稀疏问题，因此随着无标记短文本数据的到来，分类精度呈现出下降趋势。

（2）本章所提算法 SSC 在分类精度和宏平均上优于四组基准算法。原因在于 SSC 算法通过由外部语料库获取的原始词向量集合将短文本映射成 F 维特征向量，缓解了短文本的稀疏性问题，同时解决了短文本数据流存在的高维问题。然后利用丰富的无标记数据和少量的标记数据分别构建聚类器和分类器共同预测短文本数据，因此，在精度和宏平均上，SSC 高于基准算法。

（3）Snippets 数据集上原始短文本的平均长度是 News 与 Tweets 数据集的两倍以上，语义信息丰富，因此，五种算法均表现良好。REDLLA 算法是基于单模型的半监督分类模型，其在分类精度和宏平均上效果不佳， REDLLA 算法在 Snippets 数据集上低于 0.5，图中未显示。此外，从图 2 和图 3 中看出，由于短文本的稀疏性使得 TriTrain 与 SVM 在后两个数据集上的结果基本保持一致。

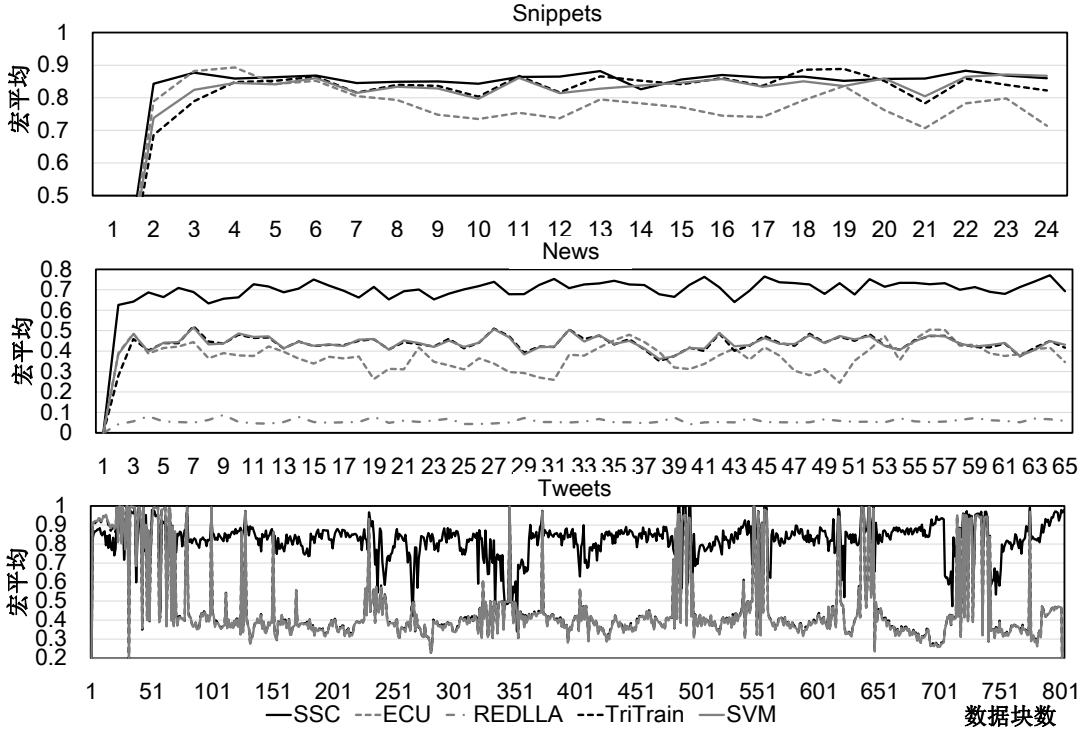

图 4.3 SSC 与基准算法在宏平均上的对比

Fig 4.3 Experimental comparison between SSC and baseline algorithms on MacroF1.

4.3.3.2 时间性能评估

本章所提算法 SSC 的时间代价主要分为分类器构建、聚类模型构建与标签传播以及分类模型对短文本的类标签预测三部分。其中标签的传播是通过 a 个已标记数据块的簇信息预测 $k-a$ 个聚类模型中聚类簇的类标签，因此其时间复杂度可以表示为 $O((a*(k-a))*r*F)$，其中 a、$k-a$ 以及 r 分别是分类器的数目，聚类器的数目以及类标签的数目，数值为定值且小，F 表示短文本的特征维度，因此 F 越小，所消耗的时间也就越小。

图 4.4 给出了 SSC 算法与基准算法在时间性能上的对比结果。由实验结果可知：

（1）SSC 算法的时间消耗明显少于 BOG 算法，原因在于 BOG 算法是基于词袋模型表示的短文本数据流分类方法，数据的高维问题导致模型的训练和预测时间更长。与其他基准算法相比，SSC 算法的时间消耗在三组数据集上均小于基准算法 ECU、REDLLA 和 TriTrain。原因在于本章所提 SSC 算法借助原始词向量集合对原始短文本数据进行降维，然后在降维后的数据上构建集成模型，而 ECU、REDLLA 和 TriTrain 算法是在原始的高维短文本数据上直接构建模型进行分类，其在分类器与聚类模型的构建上需要更久的时间，因此所花费的时间就更多。

（2）在 Snippets 和 News 数据集上，SSC 算法的时间消耗均高于 SVM 算法，但在 Tweets 数据集上要低于 SVM 算法。原因分析如下：当有标记短文本数据块到来时，SSC 算法需要不断地更新基分类器，但 SVM 算法为批处理算法，只需构建一次分类器，因此在 Snippets 和 Tweets 数据集上 SVM 算法的时间消耗低。Tweets 数据集原始特征维度过高，如表 4.1 所示，导致 SVM 算法在初始构建时就花费过高的时间，同时预测新的短文本数据块所花费的时间也高，因此本文所提算法 SSC 的时间消耗更低。

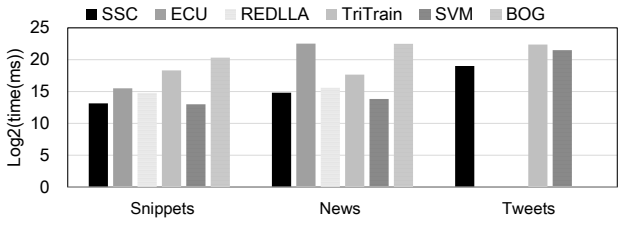

图 4.4 SSC 与基准算法的时间对比

Fig 4.4 Execution time of SSC and baseline algorithms.

4.3.3.3 概念漂移检测性能评估

Tweets 数据集存在 290 个漂移点，所提概念漂移检测算法正确检测到 239 个漂移点。由于 Tweets 数据集过大，图 4.5 仅描述了本章 SSC 算法与两个基准算法 TriTrain 和 SVM 在截取的 Tweets 的部分数据上的先序错误率结果。由图可知：与

基准算法相比，SSC 算法能够快速适应概念漂移，其先序错误率低于基准算法。主要原因分析如下：SSC 算法采用了基于聚类簇的概念漂移检测算法，当概念漂移出现时，该算法在大多数情况下都能及时发现漂移，选择未发生概念漂移的短文本数据块构建集成模型，使得该算法能快速适应概念漂移。而基准算法 TriTrain 与 SVM 为批处理算法，没有概念漂移检测机制，不适应概念漂移推文数据集。

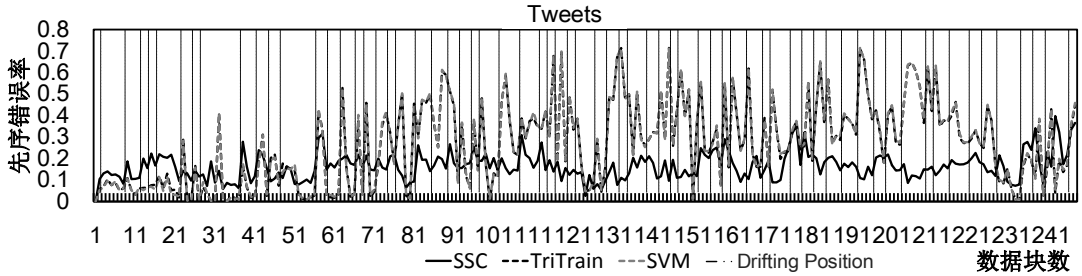

图 4.5 SSC 与基准算法在先序错误率上的对比

Fig 4.5 Experimental comparison between SSC and baseline algorithms on Prequential Error.

4.3.3.4 与第三章基于短文本扩展的 OnlineBTM 主题模型

第三章中提出了一种基于短文本扩展的 OnlineBTM 主题模型用于扩展短文本数据流，同时将短文本数据流表示成特定维度的主题。该方法挖掘外部语料库的隐藏主题并用主题下的词来扩展每条短文本数据，然后借助 OnlineBTM 主题模型将扩展后的短文本表示为 K 个主题。为对比本章所提的短文本特征向量表示方法与基于短文本扩展的 OnlineBTM 主题模型表示方法，本节用基于短文本扩展的 OnlineBTM 主题模型将短文本数据流中的每条短文本表示成一组主题，然后对主题表示的短文本数据流采用与本章一致的的半监督分类算法构建集成模型，从而预测短文本，本节将该方法称为 SSC-TBE。SSC-TBE 方法与本章所提的短文本特征向量表示方法 SSC 做对比，其分类精度对比如下：

从图 4.6 可以看出，SSC 算法在三个数据集上均高于 SSC-TBE，即基于 Word2Vec 的特征向量表示方法好于基于短文本扩展的 OnlineBTM 主题模型表示方法。原因分析如下：基于 Word2Vec 的特征向量表示方法借助外部语料库丰富的语义信息获得原始词向量集合，并用原始词向量集合中的词向量表示短文本，该方法表示的短文本具有丰富的语义信息，更有利于于标签传播。而基于短文本扩展的 OnlineBTM 主题模型表示方法则借助 LDA 主题模型挖掘外部语料库的潜在主题，并利用主题下具有代表性的词扩展原始短文本，其主题个数确定且每个主题下的代表词也是确定的，因此在扩展短文本时极易引入噪声词，例如两条不同类别下的短文本有可能会分享相同的主题，利用该主题下的表示词去扩展短文本就会影响分类结果。另外，该方法扩展短文本的词数有限，扩展后的短文本仍属于短文本，其语义信息仍然不足，将扩展后的短文本表示成主题再进行标签传播，

其效果不是很好。

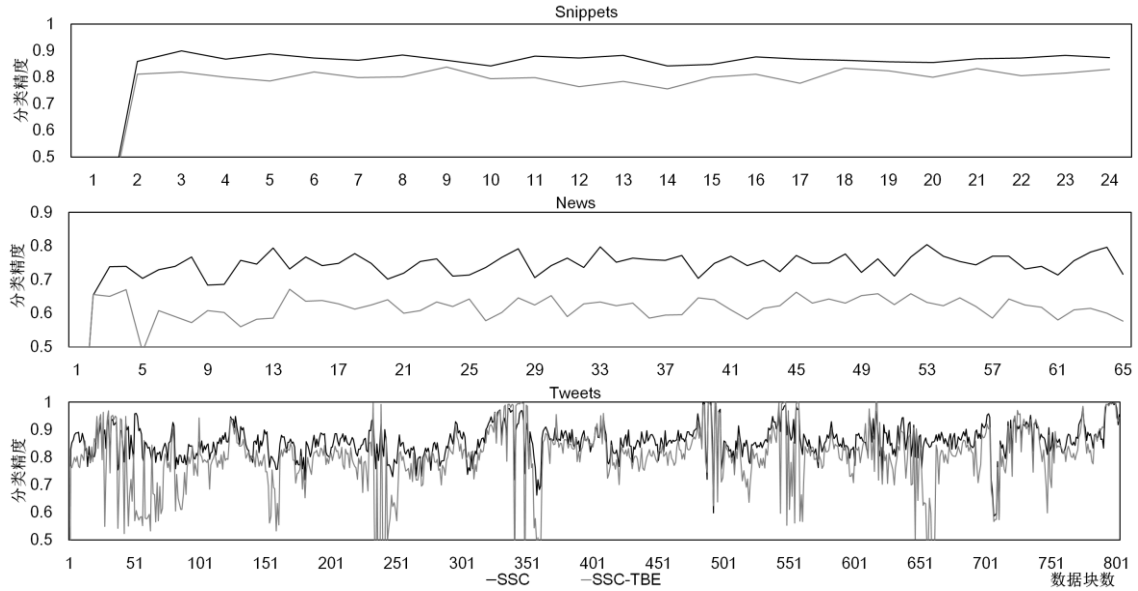

图 4.6 SSC 与 SSC-TBE 在分类精度上的对比

Fig 4.6 Experimental comparison between SSC and SSC-TBE on Accuracy.

4.4 本章小结

为解决有标记数据稀少而无标记数据丰富的问题，本章提出了一种基于标签传播的半监督短文本数据流分类算法。该算法借助 Word2Vec 工具训练来自 Wikipedia 的外部语料库获得原始向量集合，然后利用原始词向量集合将短文本数据流中每个短文本表示为 F 维向量以缓解短文本数据流的特征高维稀疏问题。同时，为充分利用丰富的无标记数据和少量的标记数据，本章针对分别针对无标记和标记数据构建聚类模型和分类器。为解决聚类模型的预测问题，本章采用一种基于簇相似度的计算方法来传递簇之间的标签信息；为适应数据流上的概念漂移，提出一种基于聚类簇的概念漂移检测算法用于集成模型。对比实验表明：所提方法具有良好的分类精度和宏平均，同时能快速适应短文本数据流上的概念漂移。

第五章 总结与展望

本章将总结针对短文本数据流分类问题开展的研究工作以及未来工作的展望。

5.1 工作总结

随着互联网技术的快速发展，海量的短文本数据不断涌现，这些短文本数据蕴含着丰富的研究价值与商业价值，如何有效的分析和挖掘这些短文本数据成为目前人们关注的重点。短文本分类作为数据分析的基础工作一直是研究者们关注的重点，然而实际短文本数据除了具有严重的语义稀疏问题外，其产生速度快且数据量大，导致传统的批处理的短文本分类算法很难适应，另外在这种流环境下的短文本数据在表示时具有严重的特征高维问题，且常常随着时间的推移出现潜在的概念漂移现象。除此之外，标记短文本数据非常稀缺，但无标记短文本数据却相对丰富的多。本文借助外部语料库缓解短文本的稀疏问题，对短文本数据流分类进行研究，主要工作如下：

（1）前两章首先介绍了短文本数据流分类的研究背景和意义，给出了现阶段短文本数据流分类存在的问题和挑战。然后，针对目前研究者们普遍关注的短文本分类问题做了相关综述，将近年来国内外研究者们对短文本分类的研究简单归为四类：基于搜索引擎的短文本分类、基于主题模型的短文本分类、基于隐藏规则和统计信息的短文本分类以及基于深度学习相关技术的短文本分类研究。另外，也简单介绍了短文本数据流分类的相关研究工作。除此之外，由于实际应用中类标签的大量缺失问题导致有监督短文本数据流分类算法通常很难被广泛应用，但是海量的无标记短文本数据具有丰富的语义信息有助于提升分类效果，因此基于半监督的分类问题也吸引了众多研究者们的关注，本文在第二章的最后给出了基于半监督的短文本分类和基于半监督的数据流分类的相关工作。

（2）本文的第三章介绍了基于文本扩展和概念漂移检测的短文本数据流分类方法，其基本思想是：首先通过分析来自 Wikipedia 的外部语料库设计了一种基于短文本扩展的在线 BTM 模型用于缓解短文本短文本数据流的特征高维稀疏问题，然后为适应数据流存在的概念漂移问题，提出一种基于主题模型的漂移检测算法，最后使用集成模型分类短文本数据同时借助漂移检测的结果来更新集成模型。

（3）针对有标记短文本数据稀少问题，本文的第四章给出了一种基于标签传播的半监督短文本数据流分类算法。为充分利用无标记短文本数据，该方法将无标记数据训练聚类器，然后通过标签传播的方法从标记数据中传播标签给聚类器的每一个聚类簇，从而使得聚类器也可以预测短文本。为了解决短文本存在的稀疏问题，该方法借助词向量工具 Word2Vec 训练来自 Wikipedia 的外部语料库获取

原始词向量集合从而扩展短文本，同时将短文本表示成固定维度的向量降低短文本数据流的特征高维问题。最后，将标记数据训练分类器联合有标记的聚类器分类新的短文本数据。

5.2 工作展望

短文本由于长度短、语义信息不足，具有严重的稀疏性，借助外部语料库的丰富语义信息扩展短文本可以缓解短文本的稀疏问题。本文所提的两种短文本数据流分类算法均借助外部语料库来缓解短文本的稀疏问题，实验结果表明这两种算法都取得了良好的分类效果。经过对借助外部语料库扩展短文本的进一步研究发现，未来可开展的工作如下：

（1）基于主题模型挖掘外部语料库的丰富语义信息从而扩展短文本是常用的扩展短文本的方法之一，该方法虽然能够有效的缓解短文本的稀疏问题，但由于主题模型的主题数是需要预先设定的，例如 LDA 和 BTM 模型，因此，针对流环境下不断变化的短文本数据来说，该方法显然具有一定的局限性。在以后的研究中，可以考虑结合不断变化的短文本数据流来调整主题数的设定。

（2）第四章所提算法中借助 Word2Vec 训练语义丰富的外部语料库获得原始词向量集合，并用这些词向量表示每条短文本中的词，然后通过加权求和方式来表示短文本。虽然该方法充分利用了外部语料库的丰富语义信息，但是忽略了短文本本身的词与词之间的顺序。在以后的短文本扩展的研究中，可以考虑短文本本身的词序的影响。

除了上述提到的有关短文本扩展的问题外，本文所用到的实验数据集都是比较小的，但在实际应用中，如何将所提的两种算法应用到海量的短文本数据流上仍然需要进一步研究。

参考文献（信息补充完整，首字母大小写统一）

[1] 2017 年微博用户发展报告[EB/OL] http://data.weibo.com/report/reportDetail?id=404.

[2] 微博开放平台-数据中心[EB/OL] http://open.weibo.com/development/datacenter.

[3] Vapnik, V. N. An overview of statistical learning theory[J]. IEEE Transactions on Neural Networks, 1999, 10(5):988-999.

[4] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.

[5] Yang Y, Liu X. A re-examination of text categorization methods[C]// International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 1999.

[6] Song G, Ye Y, Du X, et al. Short Text Classification: A Survey[J]. Journal of Multimedia, 2014, 9(5): 635-643.

[7] Bollegala D, Matsuo Y, Ishizuka M. A Web Search Engine-Based Approach to Measure Semantic Similarity between Words[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 23(7):977-990.

[8] Meng W, Lanfen L, Jing W, et al. Improving short text classification using public search engines[C]// International Symposium on Integrated Uncertainty in Knowledge Modelling & Decision Making. Springer, Berlin, Heidelberg, 2013.

[9] Tang J, Wang Y, Zheng K, et al. End-to-end Learning for Short Text Expansion[C]// The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Halifax, NS, Canada, 2017.

[10] Phan X H, Nguyen C T, Le D T, et al. A Hidden Topic-Based Framework toward Building Applications with Short Web Documents[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7):961-976.

[11] Bouaziz A, Dartiguespallez C, Célia Da Costa Pereira, et al. Short Text Classification Using Semantic Random Forest[M]// Data Warehousing and Knowledge Discovery. Springer International Publishing, 2014, 288–299.

[12] Vo D T, Ock C Y. Learning to classify short text from scientific documents using topic models with various types of knowledge[M]. Expert Systems with Applications, 2015, 42(3), 1684–1698.

[13] Zhang H, Zhong G. Improving short text classification by learning vector representations of both words and hidden topics[J]. Knowledge-Based Systems, 2016:S0950705116300193.

[14] Chen Q, Yao L, Yang J. Short text classification based on LDA topic model[C]// International Conference on Audio. 2017.

[15] Sun B, Zhao P. Feature extension for Chinese short text classification based on topical N-Grams[C]// IEEE/ACIS International Conference on Computer & Information Science. IEEE, 2017.

[16] Kim K, Chung B S, Choi Y R, et al. Language independent semantic kernels for short-text classification.[J]. Expert Systems with Applications An International Journal, 2014, 41(2):735-743.

[17] Gao L, Zhou S, Guan J. Effectively classifying short texts by structured sparse representation with dictionary filtering[J]. Information Sciences, 2015, 323:130-142.

[18] Zhang F, Yan H J, Liu M J, et al. The Research of Short Texts Classification Based on Association Rules of Lexical Items[J]. Journal of Integration Technology, 2015.

[19] Rao Y, Xie H, Li J, et al. Social emotion classification of short text via topic-level maximum entropy model[J]. Information & Management, 2016, 53(8):978-986.

[20] Li Q, Shah S, Liu X, et al. TweetSift: Tweet Topic Classification Based on Entity Knowledge Base and Topic Enhanced Word Embedding[C]// The 25th ACM International Conference on Information and Knowledge Management. ACM, 2016.

[21] Wang P, Xu B, Xu J, et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification[J]. Neurocomputing, 2016, 174(PB):806-814.

[22] Sotthisopha N, Vateekul P. Improving Short Text Classification Using Fast Semantic Expansion on Multichannel Convolutional Neural Network[C]// The 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Busan, 2018, pp. 182-187.

[23] Yan L, Zheng Y, Cao J. Few-shot learning for short text classification[J]. Multimedia Tools and Applications, 2018, 77(22):29799-29810.

[24] Zeng J, Li J, Song Y, et al. Topic Memory Networks for Short Text Classification[J]. 2018.

[25] Bouaziz A, Costa Pereira C D, Pallez C D, et al. Interactive generic learning method (IGLM): a new approach to interactive short text classification[C]// The, ACM Symposium. ACM, 2016:847-852.

[26] Ren Z, Peetz M H, Liang S, et al. Hierarchical multi-label classification of social text streams[C]// ACM, 2014:213-222.

[27] Li P P, He L, Wang H Y, et al. Learning From Short Text Streams With Topic Drifts[J]. IEEE Transactions on Cybernetics, 2017, PP(99):1-15.

[28] Cai Y H, Zhu Q, Cheng X Y. Semi-Supervised short text categorization based on Random Subspace[C]// IEEE International Conference on Computer Science & Information Technology.

2010.

[29] Chan J, Koprinska I, Poon J, et al. Semi-supervised classification using bridging [J]. International Journal on Artificial Intelligence Tools. 2008, 17(3): 415-431.

[30] Yin C, Xiang J, Zhang H, et al. A new SVM method for short text classification based on semi-supervised learning[C]// Proceedings of International Conference on Advanced Information Technology and Sensor Application. IEEE, 2016:100-103.

[31] Silva N F F D, Coletta L F S, Hruschka E R, et al. Using unsupervised information to improve semi-supervised tweet sentiment classification[J]. Information Sciences, 2016, 355-356(C): 348-365.

[32] Li X, Yan L, Qin N, et al. A novel semi-supervised short text classification algorithm based on fusion similarity[C]// Proceedings of International Conference on Intelligent.

[33] Widmann N, Verberne S. Graph-based semi-supervised learning for text classification[C]// Proceedings of ACM SIGIR Int Conference on Theory of Information Retrieval. Amsterdam: ACM, 2017:59-66.

[34] Zhang P, Zhu X, Tan J, et al. Classifier and cluster ensembles for mining concept drifting data streams[C]// Proceedings of IEEE International Conference on Data Mining. Vancouver : IEEE, 2011:1175-1180.

[35] Li P, Wu X, Hu X. Mining recurring concept drifts with limited labeled streaming data[M]. ACM, 2012.

[36] Feng Z, Wang M, Yang S, et al. Incremental semi-supervised classification of data streams via self-representative selection[J]. Applied Soft Computing, 2016, 47:389-394.

[37] Zhu L, Pang S, Ban T, et al. Incremental and decremental max-flow for online semi-supervised learning[J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(8):2115-2127.

[38] Hosseini M J, Gholipour A, Beigy H. An ensemble of cluster-based classifiers for semi-supervised classification of non-stationary data streams[M]. New York: Springer-Verlag, 2016, 46 (3) :567-597.

[39] Rakib M R H, Jankowska M, Zeh N, et al. Improving Short Text Clustering by Similarity Matrix Sparsification[C]// Proceedings of the ACM Symposium on Document Engineering, 2018, 50:1-4.

[40] Yin J H, Chao D, Liu Z K, et al. Model-based Clustering of Short Text Streams[C]// Proceedings of the 24th International Conference on Knowledge Discovery and Data Mining, London, UK, 2018, 2634-2642.

[41] Cheng X, Yan X, Lan Y, et al. BTM: Topic Modeling over Short Texts[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12):2928-2941.

[42] Zhang Y, Mao W, Zeng D. A Non-Parametric Topic Model for Short Texts Incorporating Word Coherence Knowledge[J]. 2016, 2017-2020.

[43] Tommasel A, Godoy D. Short-text feature construction and selection in social media data: a survey[J]. Artificial Intelligence Review, 2016:1-38.

[44] Tommasel A, Godoy D. A Social-aware Online Short-text Feature Selection Technique for Social Media[J]. Information Fusion, 2018, 40:1-17.

[45] Blei D M, Ng A Y, Jordan M I, et al. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2012, 3:993-1022.

[46] Wang Z, Shou L, Chen K, et al. On Summarization and Timeline Generation for Evolutionary Tweet Streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(5):1301-1315.

[47] Bifet A, Holmes G, Kirkby R B, et al. MOA: Massive Online Analysis[J]. Journal of Machine Learning Research, 2010, 11(2):1601-1604.

[48] Shalevshwartz S, Singer Y, Srebro N, et al. Pegasos: primal estimated sub-gradient solver for SVM.[C]// International Conference on Machine Learning. ACM, 2007.

[49] Pfahringer B, Holmes G, Kirkby R. New Options for Hoeffding Trees[M]// AI 2007: Advances in Artificial Intelligence. Springer Berlin Heidelberg, 2007.

[50] João Gama, Medas P , Castillo G , et al. Learning with Drift Detection[C]// Brazilian Symposium on Artificial Intelligence. Springer, Berlin, Heidelberg, 2004.

[51] Severo M, João Gama. Change Detection with Kalman Filter and CUSUM[C]// International Conference on Discovery Science. Springer-Verlag, 2006.

[52] Gama, Sebastiao, Rodrigues. On evaluating stream learning algorithms[J]. Machine Learning, 2013, 90(3):317-346.

[53] Frias-Blanco I, Campo-Avila J D, Ramos-Jimenez G, et al. Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(3):810-823.

[54] Zhou Z H, Li M. Tri-training: exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11):1529-1541.

[55] Ionescu R T, Butnaru A M. Transductive Learning with String Kernels for Cross-Domain Text Classification[J]. 2018.

[56] Fu W, Xue B, Zhang M, et al. Transductive Transfer Learning in Genetic Programming for Document Classification[C]// Asia-pacific Conference on Simulated Evolution & Learning. Springer, Cham, 2017.

[57] Beatty G, Kochis E, Bloodgood M. Impact of Batch Size on Stopping Active Learning for Text

Classification[J]. 2018.

[58] Hu R, Namee B M, Delany S J. Active learning for text classification with reusability[M]. Pergamon Press, Inc. 2016.

[59] Wang X, Mccallum A, Xing W. Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval[C]// IEEE International Conference on Data Mining. 2007.

[60] 张倩, 刘怀亮. 利用图结构进行半监督学习的短文本分类研究[J]. 图书情报工作, 2013, 57(21):126-132.

攻读硕士学位期间的学术活动及成果情况

1）参加的学术交流与科研项目

（1）国家重点研发计划项目课题六：“基于数据融合的煤矿典型动力灾害多元信息挖掘分析技术”（项目编号：2016YFC0801406）；

（2）参与国家自然科学基金“面向多源高维数据流的在线特征选择与分类方法研究”（项目编号：61673152）；

（3）参与国家自然科学基金“多标记文本数据流分类方法研究”（项目编号：61503112）。

2）发表的学术论文（含专利和软件著作权）

（1）王海燕, 胡学钢, 李培培. 基于向量表示和标签传播的半监督短文本数据流分类算法[J]. 模式识别与人工智能, 2018, 31(7):634-642.

（2）Hu X G, Wang H Y, Li P P. Online Biterm Topic Model based Short Text Stream Classification using Short Text Expansion and Concept Drifting Detection[J]. Pattern Recognition Letters, 2018.

（3）胡学钢，王海燕，李培培. 一种基于短文本扩展和概念漂移检测的短文本数据流分类方法（申请号：201710994366.9）

特别声明

本学位论文是在我的导师指导下独立完成的。在研究生学习期间，我的导师要求我坚决抵制学术不端行为。在此，我郑重声明，本论文无任何学术不端行为，如果被发现有任何学术不端行为，一切责任完全由本人承担。

学位论文作者签名：

签字日期：　　年　　月　　日