**(封面A3牛皮纸，内容与该页相同，封面内保留该页内容)**

**(页面设置：A4纸打印，版面上空2.5cm，下空2.0cm,左空2.2cm，右空2.2cm,装订线0.5cm)**

**山 东 科 技 大 学**

**毕 业 设 计（论 文）**

**说 明 书**

**设计题目：** （宋体正文小二号字）

**院 系： 专业： 班级：**

**设 计 人：**

**2020年 月 日**

|  |
| --- |
| **毕业设计任务书**  **专业 班 学生**   1. 毕业设计题目： 2. 设 计 专 题：   三、设计原始资料：      四、设计应解决下列各主要问题：        五、设计说明书应附有下列图纸：      六、命题发出日期： 年 月 日 设计应完成日期： 年 月 日  设计指导人（签章）：  教研室主任（签章）：  院 长（签章）： |

|  |
| --- |
| **特 约 评 阅 人 意 见**    评阅人（签章）：  评阅人职称：  日 期： 年 月 日 |

|  |
| --- |
| **指导教师对毕业设计的评语**    指导教师（签章）：  日 期： 年 月 日 |

|  |
| --- |
| **答辩委员会鉴定意见**    答辩成绩（等次及分数）：  **鉴定意见：**  答辩委员会主任（签章）：  答辩委员会副主任（签章）：  日 期： 年 月 日 |

**摘 要**

随着我国社会经济的迅速发展，传统能源被大量消耗……

建议(不限于)按以下段落来写：

第一段：研究背景及意义。

第二段：研究内容、方法等。

第三段：研究结论及影响。

**关键词：**(缩进两格) 分布式光伏，改进随机潮流，差分进化算法（3~5个）

近年来，随着当代科学技术的发展，各种社会媒体（如：新闻网站，社交网站，电商网站等）迅速崛起，使得网络中涌现出大量的文本数据，这些文本数据的结构与其所属领域特点往往有着较大的联系，于是跨领域的迁移学习在自然语言处理以及机器学习等领域开始得到越来越多的关注。目前传统文本分析的目标多数在同一领域，即对于某一领域的已有标签的文本数据进行机器学习，将训练出来的模型同样应用于本领域。但是在当前网络环境下，各类领域及其繁多复杂，更多的领域的数据是没有标签的，使用人工标注的方式会浪费极大的人力物力，同时，标注结果也会存在很大的主观性，因此亟需一些可实现跨领域文本标注的算法[1]。

在当前时代背景下，海量的文本信息充斥着整个互联网环境，目前传统文本分析的目标多数在同一领域，即对于某一领域的已有标签的文本数据进行机器学习，将训练出来的模型同样应用于本领域。但是在当前网络环境下，各类领域及其繁多复杂，更多的领域的数据是没有标签的，使用人工标注的方式会浪费极大的人力物力，同时，标注结果也会存在很大的主观性，因此亟需一些可实现跨领域文本标注的算法[1]。跨领域的文本特征对齐算法的相关流程主要有文本预处理，文本特征提取，特征对齐，模型检验对齐效果四个过程，下面分别对每一个过程做相关概述。

**Abstract**

With the rapid development of China’s social and economy，the traditional energy is consumed in large quantities...

尽量使用被动语态、一般现在时，若出现“论文”，用"thesis"

（英文采用Times New Roman字体）

**Keywords:** Distributed generation, improved stochastic power flow, differential evolution (专用词汇大写、除第一个关键词首字母大写外其他全部小写、标点符号是“半角逗号+空格”)

**目 录**

**1 绪 论………………………………………………………………………1**

1.1 课题的提出……………………………………………………………………………1

1.2随机潮流计算的研究现状……………………………………………………4

1.3 课题研究目的…………………………………………………………………………9

**2 随机潮流计算方法研究…………………………………………………11**

2.1 概率密度分布模型………………………………………………………………11

2.2 改进随机潮流…………………………………………………………………14

2.3 本章小结……………………………………………………………………………15

**3**

**4**

**5**

**6总结与展望………………………………………………………………70**

6.1 总结………………………………………………………………………………71

6.2 展望………………………………………………………………………………72

**参考文献……………………………………………………………………73**

**致 谢………………………………………………………………………76**

目录

[**Abstract** III](#_Toc40896303)

[1 绪 论 1](#_Toc40896304)

[1.1 课题的提出 1](#_Toc40896305)

[1.2 文本特征对齐问题的研究现状 2](#_Toc40896306)

[1.3 课题研究的目的 4](#_Toc40896307)

[2 相关概念介绍 5](#_Toc40896308)

[2.1 文本分类 5](#_Toc40896309)

[2.2 情感分类 5](#_Toc40896310)

[2.3 迁移学习 6](#_Toc40896311)

[2.3.1 基于样本的迁移学习 6](#_Toc40896312)

[2.3.2 基于基于相关性的迁移学习 6](#_Toc40896313)

[2.3.3 基于模型的迁移学习 7](#_Toc40896314)

[2.3.4 基于特征的迁移学习 7](#_Toc40896315)

[2.4 文本特征与特征选择 7](#_Toc40896316)

[2.4.1 卡方统计(CHI)法 7](#_Toc40896317)

[2.4.2 信息增益(Information Gain, IG)法 8](#_Toc40896318)

[2.4.3 文档频率(Document Frequency, DF)统计法 8](#_Toc40896319)

[2.4.4 互信息(Mutual Information, MI)方法 9](#_Toc40896320)

[2.5 文本特征对齐 9](#_Toc40896321)

[2.6 分类器 10](#_Toc40896322)

[2.6.1 支持向量机(SVM) 10](#_Toc40896323)

[2.6.2 朴素贝叶斯(Naïve Bayesian Classifier) 11](#_Toc40896324)

[2.6.3 决策树(Decision Tree) 12](#_Toc40896325)

[2.6.4 k近邻(k-nearest neighbor) 14](#_Toc40896326)

[2.7 本章小结 16](#_Toc40896327)

[3 文本特征对齐任务及数据预处理 16](#_Toc40896328)

[3.1 文本特征对齐任务 16](#_Toc40896329)

[3.1.1 任务介绍 16](#_Toc40896330)

[参考文献 18](#_Toc40896331)

# 1 绪 论

## 1.1 课题的提出

近年来，随着当代科学技术的发展，各种社会媒体（如：新闻网站，社交网站，电商网站等）迅速崛起，使得人们可以通过越来越多的方式在网络上发表自己的观点，致使当前国内外网络环境中涌现出大量的文本数据。这些文本数据的结构与其所属领域特点往往有着较大的联系，有些词汇甚至可以成为某些领域的代名词，例如：当我们看到“好评”、“物美价廉”等词语时，我们会想到这可能是对于某一商品的评价；当我们看到“吓人”、“振奋人心”等词语时，我们会想到这可能是对于某一消息或新闻的评价。于是一些情感分类算法在机器学习领域应运而生，但是这些情感分类算法的分析目标多数是在同一领域，即对于某一领域的已有标签的文本数据进行机器学习，将训练出来的模型同样应用于本领域，从而完成对未分类的文本的分类任务。但是在当前网络环境下，各类领域及其繁多复杂，而更多的领域的文本数据是没有标签的，同时使用人工标注的方式会浪费极大的人力物力，并且标注结果也会存在很大的主观性，导致标注结果的准确性极大的降低，因此在文本分析中亟需一些可实现跨领域文本标注的算法[1]，来对已有文本进行情感分类。

笔者认为，目前国内经济发展水平呈上升态势，人民群众正在不断谋取更高的物质生活水平与精神生活水平，于是网上购物与新闻知识获取的向往日趋强烈，使得电商领域与新闻领域成为全体网民乐于聚集之地，这两个领域所涵盖的人群数量大且范围广，因此在两个领域中留下了的文本信息较为全面，对这两个领域做迁移学习将有更大的研究价值，例如国家可以通过文本情感分析结果对社会舆论信息进行监控；对于包括新闻网站在内的大多数网站来说，可以对用户评论内容进行关键词提取进而对不符合要求的敏感词进行屏蔽；电商平台可以利用用户评论信息进行产品优化或修改营销计划；一些音视频媒体也可以根据用户对某些影视剧或某些歌曲的评论情况对其他用户进行择优推荐等等。目前面向迁移学习的文本情感分类并没有成文的步骤，笔者在阅读大量文献之后，总结得出做文本情感分类方面的迁移学习应当有如下4个步骤：1、文本预处理，2、文本特征提取，3、特征对齐，4、模型检验。显然，第2与第3步应当是整个过程的重中之重，对于基于特征的迁移学习而言，源域与目标域之间的枢纽特征的匹配对于在两个领域中进行机器学习的效果是至关重要的，因此研究好如何将两个领域中的非枢纽特征更好地转化为枢纽特征，以及如何将两个领域中的枢纽特征做最大程度上的对齐，对于提升迁移学习的准确率是至关重要的，这是选取本课题的重要意义之所在。

本文中，笔者首先对当今时代下跨领域文本特征对齐算法进行全面剖析，而后选取对本课题最为契合的算法，以电商领域为源域，以新闻领域为目标域，完成两个领域之间的迁移学习。

## 1.2 文本特征对齐问题的研究现状

目前网络环境中，各类文本信息有着多种多样的形式，它们所表达的情感倾向也相差甚大，例如在电商领域与新闻领域中，用户们会发表许多的关于产品或者时政的评论及意见，这些评论及意见通常含有用户的情感极性。Turney[2]等人在2002年第一次将文本情感分类的任务当做学术问题公开提出，之后再行业内引发了众多学者的研究。

目前常用的基于统计的特征选择方法有信息增益方法(IG)、互信息方法(MI)、卡方统计(CHI)法、文档频率方法(DF)等[3-5]。2010年，在总结已有特征选择方法的基础上，杨凯峰[6]等一种新的基于文档频率(DF)的特征选择方法，此方法主要用于中文文本分类，需要先对每篇文档进行分词处理。他们使用 ICTCLAS(Institute of Computing Technology,Chinese Lexical Analysis System)系统完成文档的分词及词性标注，使用特征词在一个类别中出现的文档数来表示这个特征词与该类别的相关度，他们使用复旦大学整理的语料库进行分类实验，结果显示，在进行特征选择时，补充部分高词频特征词可以提高分类的召回率和准确率。在向量空间模型中，将文本表示单元（特征词）转换为向量通常使用的权重计算方法是TF-IDF(term frequency-inverse document frequency)方法，在一定程度上，这个方法能有效地表示一个特征词在文本中区分文本属性的重要程度，但其理论依据存在不足。2013年路永和[7]等提出了改进TF-IDF算法的文本特征项权值计算方法，改进的核心思想是：每个特征词对每个类的区分能力不同，其重要性用特征选择评估函数反映，同时该词集中分布的类就是其所属的类，即这个特征词要具有这个类的特征性。为避免特征词只集中出现在某篇文档中，要求该特征词在该类别文档中应均匀分布。随后，曹鲁慧[8]等提出一种基于深度学习的中文文本特征提取与分类方法，她们使用卷积神经网络(CNN)进行文本分类，将每个样本用128维的向量进行表示，大大减少特征的维度，加快分类器的训练速度，进而提高分类的准确率。卢晨阳[9]等提出一种基于语义结构的迁移学习文本特征对齐算法，他们使用wordnet模型完成领域独立特征的对齐。他们考虑到，传统的枢纽特征抽取和对齐一般只考虑到句子的结构和词性等信息，而忽略词语特征和类别信息存在一定的相关性，因此在抽取特征时应考虑类别影响。以情感分类为例，在抽取特征训练模型时，首先将源域和目标域的正类样本(积极)和负类样本(消极)分开，将不同词性的词语抽取出来，其中，在情感分类中起主要影响的是名词、动词、形容词、副词，然后对正类样本和负类样本分别抽取名词、动词、形容词和副词，原始的源域和目标域的2份数据变成根据样本类别和词性划分的16 份数据。最后 使用16 份数据作为词向量模型GloVe 模型的输入来进行训练。

另外，学者们对文本特征对齐算法的研究也在不断深入，2006年，John Blitzer[10]等提出一种结构一致学习算法(Structure Correspondence Learning, SCL)，这种算法可以自动归纳来自不同领域的特征之间的对应关系，SCL的核心思想是通过建立不同领域特征与枢纽特征的相关性来识别特征之间的对应关系。SCL的第一步是在两个域的未标记数据上定义一组枢纽特征。然后，使用这些枢纽特征来学习从两个域的原始特征空间到共享的低维实值特征空间的映射。在这个新的空间中，如果内积很高，则表示对齐效果较好。在有监督的训练任务中，同时使用源域中的转换特征和原始特征。在有监督的测试任务中，则同时使用目标域的转换特征和原始特征。如果学习到的映射比较好，那么在源域上学习到的分类器也将在目标域上有效。2013年，孟佳娜[11]等提出一种基于特征变换的跨领域产品评论倾向性分析方法，她们使用一种特征变换算法(Feature Transform Algorithm,FTrA)，首先计算源领域和目标领域的领域独立词，然后，求得源领域的领域依赖词与每个领域独立词之间的关联度值，与某个领域独立词关联度最高的特征，与其相关性越高。同时，求得目标领域的领域依赖词与每个领域独立词之间的关联度值，与某个领域独立词关联度值越高的特征，与其相关性越高。在此基础上，Sinno Jialin Pan[12]等提出了一种谱特征对齐 (Spectral Feature Alignment,SFA) 算法。SFA以一些领域无关词为桥梁，构造了一个二部图，来模拟领域特定词和领域无关词之间的共现关系。其思想是：如果两个特定于域的词与图中更常见的独立于域的词有关联，它们往往会以更高的概率对齐。类似地，如果两个独立于域的词与图中更常见的特定于域的词有连接，则它们倾向于以更高的概率排列在一起。他们采用基于图谱理论[13]的谱聚类算法，在二部图的基础上，将特定于领域的和与领域无关的词联合成一组特征聚类。这样，集群就可以用来减少两个域的特定于域的词之间的不匹配。最后，他们用这些聚类来表示所有的数据实例，并基于新的表示训练情感分类器。与结构一致学习(SCL)[10]等跨域情感分类算法不同，他们提出的SFA算法可以通过在二部图上的联合对齐，充分利用领域无关词和领域特定词之间的关系，学习更紧凑、更有意义的表示图的底层。通过实验表明，在跨领域情感分类的准确性方面，SFA确实比SCL等具有更好的性能。但是Yuhong Zhang[14]等认为，SCL与SFA这类的算法尝试提取出域无关主题词来构造子空间，并在子空间中训练分类器，这样的话，目标域中的域相关特征会被映射为源域的特征，然而在实际应用当中，这些方法无法解决在不同域中具有不同情感的特征词的问题，因为一个主题仅仅能够表达出一种情感，这些特征就会在源域中被错误地训练。针对这一问题，Yuhong Zhang等人提出一种面向跨领域情感分类的公共子空间构造算法[14]，此算法的目标是构造出一个更加精确的公共子空间，然后就可以在这个子空间中为目标域训练出分类器。对于有标签的源域S和无标签的目标域T，他们使用分类比例差[15]( Categorical Proportional Difference, CPD)算法计算源域中特征词的情感倾向，然后基于共现关系预测目标域中这些特征的情感取向，根据不同域之间情感取向的一致性构造公共子空间，于是域相关特征将会被过滤掉，最终使所有的域相关特征都会被投影到公共子空间中来解决域失配问题。跨领域的文本特征对齐是做文本情感分析的迁移学习过程中的重要一步，对与文本特征对齐技术与算法方面的研究，时至今日，仍吸引着众多的学者来专研，这也是笔者以此为课题的原因之一。

## 1.3 课题研究的目的

正文模板中所有正文数字及字母应使用新罗马字体。

# 2 相关概念介绍

## 2.1 文本分类

文本分类[16]是指在预定义好的分类体系下，依据文本的特征（例如属性及内容），将给定的文本数据与一个或者多个类别相关联的过程。文本分类方面的研究会涉及到文本内容理解和模式分类等若干个自然语言处理和模式识别方面的问题，一个文本分类系统不仅是一个自然语言处理系统，也是一个模式识别系统，系统的输入内容是待分类处理的文本，系统的输出是与该文本所关联的类别。

总体来说，在国内外对于文本分类方面的研究共经历的下面4个主要的发展阶段：

1、第一发展阶段（1958年-1964年）：这一阶段的主要贡献是对文本自动分类的可行性进行了研究。

2、第二发展阶段（1965年-1974年）：这一阶段的主要贡献是进行文本自动分类的实验研究。

3、第三发展阶段（1975年-1989年）：在这一阶段中，成型了一些文本自动分类方法， 学者们的研究重点转向对已有算法的深入探索与改进。

4、第四发展阶段（1990年至今）：这一阶段为面向全网的文本自动分类方面的研究。

## 2.2 情感分类

情感(Sentiment)，是指人们对于某件事物产生的内在的主观感受。情感分类（有时被称为倾向性分析）是指根据文本内容所表达出来的含义的不同，将不同的文本划分成包括两种（二分类）或多种类型（多元分类），是对文本发表人员观点或态度的划分。

情感分类问题不同于其他的一些分类问题，它既包含传统分类中的共性问题，同时它又有自己的独特性，例如，有些作者在进行情感信息表述时，他所表达出来的情感倾向可能会比较模糊，从而导致情感信息有着较大的多义性，也就是极性不明显的问题。针对这一问题，研究者们通过不懈的努力，在做了大量研究之后，提出了众多的情感分类方法，这些方法可以按照不同的归类方式划分:按所给文本中的情感特点分类，按机器学习方法划分。按所给文本中的情感特点的不同，情感分类研究可划分为数据不平衡性问题研究和领域相关性研究两类。若根据机器学习方法的不同，可以将文本情感分类分为有监督学习方法（训练数据既有特征(feature)又有标签(label)）、半监督学习方法（训练集同时包含有标记样本数据和未标记样本数据）和无监督学习（训练样本的标记信息未知）。

## 2.3 迁移学习

如今的许多机器学习方法要求训练数据和测试数据必须从同一特征空间（即本文中所说的域）中获得，并且需要具有相同或相近的分布。当分布情况发生较大的改变时，大多数的统计模型需要使用新收集的训练样本重新构建。但是在许多现实的应用中，重新收集所需要的训练数据来对模型进行重建，是需要花费很大代价的。在这种情况下，迁移学习(Transfer Learning)[18]的优势就显而易见了。首先，当我们发现待处理空间与当前用来构建模型的空间有着较大的差别时，我们不必丢弃现有工作而从新开始，使用迁移学习完全可以帮助我们复用当前的研究成果。其次，迁移学习有着强大的通用性和时效性，无论是在无监督学习、半监督学习还是有监督学习中，迁移学习都有着广泛的应用。如今，迁移学习可分为四种类型：基于样本的迁移学习、基于相关性的迁移学习、基于模型的迁移学习和基于特征的迁移学习。

2.3.1 基于样本的迁移学习

基于样本的迁移学习(Instance-based Transfer Learning)在使用中需要先对样本进行一种分类操作，将所有的样本按重要程度进行划分，用不同的权重值来代表不同的重要程度，另外还需要找到源域与目标域中共有的或相类似的数据，对这些数据放大之后，再到目标域中进行应用。这样的方法优点是方法简单，实现容易。缺点在于权重的选择与相似度的度量依赖经验，且源域与目标域的数据分布往往不同。

2.3.2 基于基于相关性的迁移学习

基于相关性的迁移学习(Relation-based Transfer Learning)适合应用在当两个与之间的关系比较相似的时候，例如以生物病毒的传播模型来预测计算机病毒的传播情况。当两个域相似的时候，它们之间会有共享某种相似关系，典型的方法是Mapping方法。

2.3.3 基于模型的迁移学习

基于模型的迁移学习(Model-based Transfer Learning)是一种比较直接的迁移学习方法，对于在源域中通过大量测试数据训练出来的模型，将它直接应用到目标域中，只需要对模型的参数进行微调，就可以达到一个比较好的训练效果，这样的方法的有点是可以充分利用两个领域模型之间存在的相似性，但参数调整过程较为繁琐，通常难以得到一个很好的参数组合。

2.3.4 基于特征的迁移学习

基于特征的迁移学习(Feature-based Transfer Learning)要求源域和目标域中含有一些共同的交叉特征，这时我们可以通过特征变换，将源域和目标域的特征变换到相同特征空间，使得该特征空间中源域数据与目标域数据具有相同的数据分布，然后在此基础上进行传统的机器学习。此类迁移学习方法的优点是对大多数领域适用，且效果较好，缺点在于难于求解，容易发生过适配的问题。需要注意的的是基于特征的迁移学习方法和基于实例的迁移学习方法的不同是基于特征的迁移学习需要进行特征变换来使得源域和目标域数据变换到同一特征空间，而基于实例的迁移学习只是从实际数据中进行选择来得到与目标域相似的部分数据，然后直接学习。

## 2.4 文本特征与特征选择

文本特征是文本中与文本表达的意图相匹配或能够代表文本意图的字、词、短语或其他元素，因此，在做自然语言处理或文本方面机器学习时，文本特征提取的准确与否将直接影响到模型的最终效果。这时就需要一个良好的特征选择算法，目前已有的特征提取方法比较多，常见的有：卡方统计(CHI)法、信息增益(Information Gain, IG)法、文档频率(Document Frequency, DF)统计法和互信息(Mutual Information, MI)方法等，下面对这四种方法做简单介绍。

2.4.1 卡方统计(CHI)法

该方法使用一个卡方值来衡量某一特征t与某一类别c之间的关联度，卡方值的计算公式（2.1）如下：

其中，N表示训练集中的文本总量，A表示包含特征t且在类别c中出现的文档频率，B表示包含特征t且不在类别c中出现的文档频率，C表示不包含特征t且在类别c中出现的文档频率，D表示不包含特征t且不在类别c中出现的文档频率。对于计算结果卡方，其值越高，表示某特征项与该类别之间的关联度越高，以此来达到特征选取的效果。

2.4.2 信息增益(Information Gain, IG)法

在了解信息增益法之前，需要先了解熵(Entropy)的概念，对于一给定的文本信息，熵指的是这一文本所包含信息量的多少。熵的计算表达式如公式（2.2）所示：

信息增益(IG)法就是根据某一特征在样本中所能提供熵值的多少来衡量此特征的重要程度，然后根据这个重要程度来决定此特征的去留。对某个特征来说，它的信息增益是指在样本中当不考虑此特征时的信息熵与考虑此特征值时的信息熵之间的差值。用公式表示如下：

其中，代表在样本中，类型的文本的出现概率，代表含有特征的文本的出现概率，代表不含有特征的文本的出现概率，代表某文本在包含特征的条件下又属于类型的条件概率，代表某文本在不包含特征的条件下又属于类型的条件概率，若用NUM来表示所有文本的类别数，则变量j的变化范围是。

2.4.3 文档频率(Document Frequency, DF)统计法

从训练集中统计出包含某些特征的文档的出现频率，然后需要对所有的文档频率设定两个文档频率阈值，分别为最高阈值与最低阈值，当某一个特征的文档频率超过这个最高阈值时，我们需要将该特征删除，原因是这种特征区分度过低；当某一个特征的文档频率低于这个最低阈值时，我们也将该特征删除，原因是这种特征往往不具备代表性。DF方法可以去掉一部分无关特征值，因此可以有效地降低模型的计算复杂度，并且能够提高文本分类的精准度，且算法比较简单，实现方便。

2.4.4 互信息(Mutual Information, MI)方法

此方法的算法原理为：当一个特征与类型之间的关联程度越大时，其互信息值越大。互信息值的计算公式如下式（2.4）所示：

式（2.4）中的所有字母所代表的的含义与前面三个方法中使用的字母含义一致。从公式可知，当与无关时，与的值相等，最终互信息计算结果为0，此时特征与类型之间的关联程度最低。

## 2.5 文本特征对齐

下面介绍在特征对齐过程中应用到的一些概念。

领域：对某一类别下事物的一种抽象的集合表示，在本文所讨论的此类文本分类任务中，领域即为两个行业中文本数据的集合。

源领域：有标签的文本所在的领域，在本课题中即电商领域。

目标领域：无标签的文本所在的领域，在本课题中即新闻领域。

枢纽特征：在源域与目标域中共同出现且出现频率较高的特征，这些特征可以作为在两个领域中做特征对齐时的桥梁，用于将两个领域中的相似枢纽特征联系起来。

非枢纽特征：除了枢纽特征之外的其他特征即为非枢纽特征。

相似枢纽特征：与枢纽特征相似度较高的非枢纽特征即为相似枢纽特征。

跨领域的文本特征对齐：通过一些算法，将两个领域中的非枢纽特征与枢纽特征进行关联，从而使在源领域中训练得到的文本分类器能够在目标领域中得以应用进而完成迁移学习，可以用如下四种算法来完成文本特征对齐任务：结构一致学习(SCL)算法，谱特征对齐 (SFA) 算法，特征集成及样本选择(SS-FA)算法[19]，Word2vec算法。

## 2.6 分类器

文本情感分类问题归根结底也是一个分类问题，分类问题的常用算法在文本情感分类中同样适用，所谓分类就是根据对象的特征将对象分配到相应的类别中去。目前的分类器或分类算法有很多，这里对常用的几个算法进行简单介绍：支持向量机(Support Vector Machaines, SVM)分类器，朴素贝叶斯(Naïve Bayesian Classifier)分类器，决策树(Decision Tree)分类器，k近邻(k-nearest neighbor)算法。

2.6.1 支持向量机(SVM)

支持向量机(SVM)分类器：目标是在训练集数据中寻找一个决策平面(decision surface)，这个超平面可以将训练集中的数据做最大程度上无误差的划分。支持向量机主要用于解决二元分类问题分。SVM的基本概念有如下三点：

线性分类：

二分类的问题通常使用一个线性函数（n是训练集样例的维数，R是实数集），对于函数的输出，我们可以规定：当函数f的输出值时，认为X的分类结果为负类，当函数f的输出值时，认为X的分类结果为正类。的通用形式如下式（2.5）所示。

f(x)即为所要寻找的超平面，这个平面可以将所有的输入（也就是x）分成正类与负类两半，如下图2.1所示。

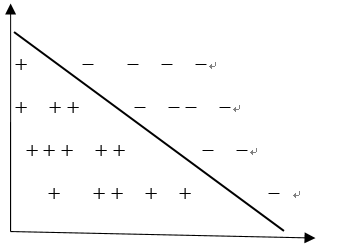


图2.1 f(x)所形成的超平面示意图

对于多分类问题，那么f(x)的值域应当是，将二分类中的线性模型向m类中转型就需要对每个类都关联一个相应的与，那么新的决策函数如下式（2.6）所示，这样可以为每一个类创建一个超平面。

线性不可分：

若当前问题是一个非线性问题，那么可以通过映射的方式，将样本集x映射到另一个空间中，然后在另一个空间中使用前述的线性分类，我们使用表示从空间X到空间F的映射，那么新的决策函数如下式（2.7）所示：

构造核(kernel)函数

核函数要求，对于, 满足

这里的是从空间X到空间F的映射。

在引入了核函数后，决策函数就可以用下式（2.9）表示，

此时，此算法的关键就在于核函数的建立，核函数要求适用的特征空间必须是对称的，也就是：

另外，核函数还必须满足Mercer定理：X的任意子集对应的矩阵必须是半正定的，综上是核函数的充要条件为：

经过以上介绍，可知，SVM算法最适用于二分类问题，且更适用于文本分类，目前文本分类方面的研究学者多数是使用的本算法来完成的文本情感分类的任务，因此在本文中，本人也是用这种算法来完成迁移学习任务。

2.6.2 朴素贝叶斯(Naïve Bayesian Classifier)

朴素贝叶斯算法的核心思想是将特征和类别发生的联合概率作为给该文本分类的概率，因此要使用朴素贝叶斯算法首先需要对条件概率的相关内容有所了解，在条件概率中，在A事件发生的条件下B事件发生的概率记做P(B|A)，并且P(B|A)的计算公式如下式（2.12）所示：

在朴素贝叶斯算法中，假设训练集有d组数据，共有m个类别，其计算公式如下式（2.13）所示：

此公式表示，对于某一测试样例，分别算出此样例发生的条件下，属于类别的概率，去概率最大值对应的那个类别，即为朴素贝叶斯算法分类类别。

对于的计算，当x的各个属性之间相互独立时，若x有k个属性，可通过下式（2.14）计算得出

其中的计算方式如下：若是彼此无关的属性序列，对于训练集中的数据对象，属性下的属性值有n个，在类别下，相应的属性值有m个，那么的计算公式为（2.15）：

对于（2.13）式中的与，分别在训练集中统计所有样本中相应的事件发生的概率即可。经过以上计算方法，以概率高低来进行分类，这样的分类结果能够比较贴合真实情况，但是，经过以上算法流程的介绍，可以看出，此算法比较适应于多属性，多分类任务，对于本文所讨论的面向文本处理的二分类任务，作者认为并不适合。

2.6.3 决策树(Decision Tree)

决策树(Decision Tree)是一类常见的机器学习方法。顾名思义，决策树就一种基于树的结构来进行决策的算法，样本集中的全部数据从树的根节点开始判断，没经过一个节点进行一次判断，根据判断结果将该样例划分到相应的子树中去，重复上述过程，直至到达叶子节点，这时每一个叶子节点都有一个唯一对应的分类类别，所有叶子节点中的分类类别通常是存在重复的。决策树可以是二叉树也可以是多叉树，关键在于每一个属性存在几类属性值，以二分类任务为例，决策过程如图2.2所示。

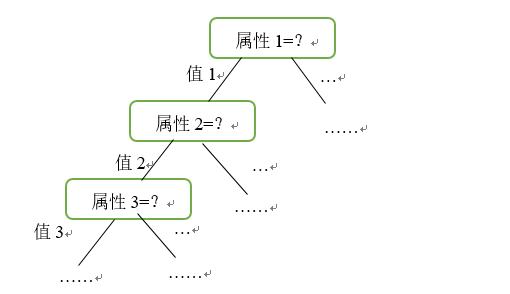


图2.2 二分类决策树示意图

一棵决策树的形成过程包括两个阶段：构造与修剪。构造就是选择何种属性作为节点，确定每个节点的划分，最终生成一颗完整的决策树的过程；修剪就是对决策树进行“瘦身”的过程，在这个过程中将一些多余的判断结果删除，使我们构造的决策树能够更加简洁，防止“过拟合”的情况发生。

“过拟合”是指模型的训练结果太好，以至于在实际应用的过程中，会存在“死板”的情况，导致分类错误，造成“过拟合”现象的原因是通常是因为训练集中样本数量较少，如果决策树选择的属性过多，构造出来的决策树一定能够极好地把训练集中的样本分类，但是这样就会把训练集中某一些数据的特点当成是所有数据的特点，这就使得这个决策树在真实的数据分类中会出现错误，也就是模型的“泛化能力”差。这里的“泛化能力”是指一个模型应对训练集中未存在的情况的能力，也就是“举一反三”的能力，当我们在训练模型时如果太依赖于训练集中的数据，那么就可能出现前述“过拟合”现象中存在的问题，也就导致了“泛化能力”差的现象。有“过拟合”自然就会存在“欠拟合”的情况，“欠拟合”是指模型的训练效果未达到最佳的情况，某些情况可能无法进行分类，这也会导致分类错误情况的发生。

评价一个决策树模型的指标有纯度和信息熵。如前2.4.2节所述，信息熵用于表示信息的不确定度，当不确定性越大时，信息熵也就越大，另外，决策树的构造过程可以看做是一个净化过程，也就是将“纯度”提高的过程，显然，信息熵与纯度的关系是：信息熵越大，纯度越低，当样本集中的所有样本均匀混合时，信息熵最大，纯度最低。

2.6.4 k近邻(k-nearest neighbor)

顾名思义，k近邻是指k个距离最近的邻居，那么k近邻算法原理就是，对于一个待分类的样本，查找出距离这个待分类样本最近的k个已分类样本，在这k个已分类样本中，哪一个类别的样本拥有得最多，那么这个样本就会被分到该类别中去。如下图2.3所示，以正方形代表一个待分类的样本，以三角形和圆圈代表两中已分类的样本，以下图表示以个二分类任务，当k取3时，在距离正方形最近的3个样本中，三角形最多，所以该样本将被分派到三角形所在的类别。

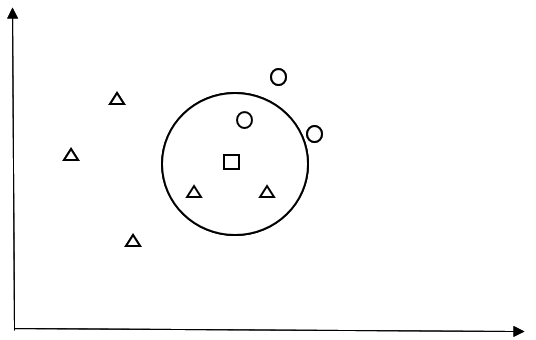


图2.3 当k=3时，knn算法的决策结果

如图2.3所示，当k值取5时，在距离正方形最近的5个样本中，圆圈最多，所以该样本将被分派到圆圈所在的类别。

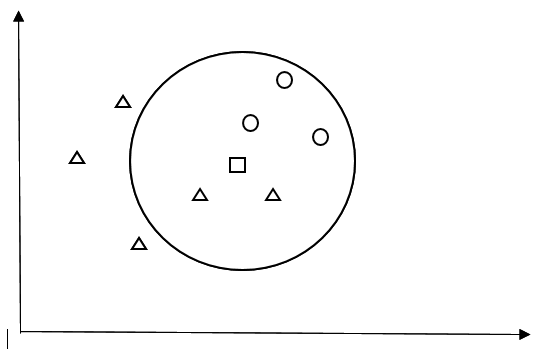


图2.3 当k=5时，knn算法的决策结果

经过两张图的对比可以发现，k值是一个影响决策结果的关键因素，k值的选择直接关系到分类器的运算结果，本算法另一个关键点在于点与点之间的距离怎么算，下面着重介绍这两点。

k值选取

k值的选取可以使用交叉验证法，将样本集按一定比例划分成训练集和测试集之后，先从一个比较小的k值开始，分别计算某一个k值下验证集合的方差，可以作出一个方差值随k值变化的折线图，方差值越小说明此时的k值的分类效果最佳，所绘制的折线图应当如下图2.4所示的形状。

图2.4 分类方差随k值变化示意图

此折线图中，方差值应当会随着k值的增大而先减小后增大，因为起初k值比较小，可用于判别的样本数量太少，所以分类误差会比较大，随着k值的增大，方差会逐渐减少，但到达某一k值之后，由于k越来越接近于样本集中的样本总数，分类效果会越来越差，错误率会再次提升，所以对于作出的折线图，我们应当取最低点对应的那个k值。

距离计算

度量空间中点之间距离的方式有多种，常见的有曼哈顿距离、欧氏距离等。

二者的计算公式如下：

曼哈顿距离：

欧氏距离：

在knn算法中，我们常用欧氏距离来计算两个样本之间的距离。

## 2.7 本章小结

本章对本文研究方向中使用到的一些重要概念进行了详细介绍，首先对文本分类、情感分类、迁移学习此类的大型概念进行描述，其中对于一些在机器学习，文本分类任务中常用的算法，文本特征提取算法均进行了细致解读，以上均是作者本人在阅读大量文献之后，根据自身体会而总结出的算法机理。由于本课题任务是实现跨领域的文本特征对齐，因此在2.4节对于在文本特征提取中使用频率较高的四个算法进行了细致解读，因为研究的目的是实现跨领域的迁移学习，因此对机器学习中作者认为比较主流的分类器算法进行了阐述，以便为后面章节的研究工作打下基础。

# 3 文本特征对齐任务及数据预处理

## 3.1 文本特征对齐任务

3.1.1 任务介绍

2**随机潮流计算研究**

**2.1 概率密度分布模型**

参考文献

专著：[序号] 编著者. 书名[M]. 出版地: 出版社, 年代, 起止页码

期刊论文：[序号] 作者. 论文题目[J]. 期刊名称, 年度, 卷(期): 起止页码

学位论文：[序号] 作者. 学位论文名称[D], 发表地: 学位授予单位, 年度

会议内容：[序号] 作者. 论文题目[C]. 会议名称, 出版年, 起止页码

参考文献例子如下：

[1] 王锡凡. 现代电力系统分析[M]. 北京: 科学出版社, 2003, 14-15.

[2] 丁明, 李生虎, 黄凯. 基于蒙特卡罗模拟的概率潮流计算[J]. 电网技术, 2001, 10(11): 10-14.

[3] 郭靖. 原对偶内点法及分枝定界法在无功优化中的应用[D]. 济南: 山东大学, 2004.

[4] Li W, Yang G, Zhou Z, et al. Impact of the distributed photovoltaic on the current protection of 10kV distribution network[C]. Power and Energy Engineering Conference, 2014, 1-4.

参考文献格式要求

1、参考文献格式需要统一，其中**所有标点符号**为半角，而且里边的标点符号后边加空格.

2、参考文献不要大量引用学位论文[D], 多引用近5年的成果论文, 尤其是16 、17年的论文

3、作者先写姓后写名，名字简写后面不加点。例如 Huang Q X,

4、参考文献的类型一定检查仔细，[J]、[C]等

5、参考文献作者多于三人时，列出前三位作者，后面几位用“等”字(英文使用et al)代替；如“张三, 李四, 王五, 赵六.”写成 “张三, 李四, 王五, 等.”

1. 魏晓聪, 林鸿飞. 面向迁移学习的文本特征对齐算法[J]. 计算机工程, 2017, 43(2): 215-219, 226.
2. Turney P D, Littman M L. Measuring praise and criticism: Inference of semantic orientation from association[J]. ACM Transactions on Information Systems(TOIS), 2003, 21(4): 315-346.
3. Galavotti L, Sebastiani F, Simi M. Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization[C]// Research & Advanced Technology for Digital Libraries, European Conference, Ecdl, Lisbon, Portugal, September. Springer-Verlag, 2000.
4. Mladenic D, Brank J, Grobelnik M, et al. Feature selection using linear classifier weights: interaction with classification models[C]// SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004. ACM, 2004.
5. 林森, 唐发根. 基于Log似然比的特征选择算法[J]. 计算机工程, 2009(19): 62-64+67.
6. 杨凯峰, 张毅坤, 李燕. Feature Selection Method Based on Document Frequency%基于文档频率的特征选择方法[J]. 计算机工程, 2010, 036(017): 33-35, 38.
7. 路永和, 李焰锋. 改进TF-IDF算法的文本特征项权值计算方法[J]. 图书情报工作, 2013(03): 92-97.
8. 曹鲁慧, 邓玉香, 陈通等. 一种基于深度学习的中文文本特征提取与分类方法[J]. 山东科学, 2019, 32(06): 106-111.
9. 卢晨阳, 康雁, 杨成荣等. 基于语义结构的迁移学习文本特征对齐算法[J]. 计算机工程, 2019, 45(05): 116-121.
10. Blitzer J, Mcdonald R T, Pereira F. Domain Adaptation with Structural Correspondence Learning[C]// EMNLP 2007, Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 22-23 July 2006, Sydney, Australia. 2006.
11. 孟佳娜, 段晓东, 杨亮. 基于特征变换的跨领域产品评论倾向性分析[J]. 计算机工程, 2013, 39(10): 167-171.
12. PAN Jialin, NI Xiaochuan, SUN Jiantao, et al. Crossdomain sentiment classification via spectral feature alignment[C]// Proceedings of the 19th International Conference on World Wide Web. New York, USA: ACM Press, 2010:751-760.
13. Chung F R K. Spectral graph theory [M]. American Mathematical Soc, 1997.
14. Zhang Y, Xu X, Hu X. A Common Subspace Construction Method in Cross-Domain Sentiment Classification[C]// International Conference on Electronic Science & Automation Control. 2015.
15. M.Simeon, and R.Hilderman, “Categorical proportional difference: A feature selection method for text categorization,” The Australasian Data Mining Conference. pp. 201–208, 2008.
16. 宗成庆. 统计自然语言处理[M]. 北京: 清华大学出版社, 2013, 416-417.
17. 肖明. WWW科技信息资源自动标引的理论与实践研究[D]. 中国科学院文献情报中心, 2001, 45-48.
18. Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on knowledge and data engineering, 2010, 22(10): 1345-1359.
19. Xia R, Zong C, Hu X, et al. Feature Ensemble Plus Sample Selection: Domain Adaptation for Sentiment Classification[J]. IEEE Intelligent Systems, 2013, 28(3): 10-18.
20. Zhang S, Liu H, Yang L, et al. A Cross-Domain Sentiment Classification Method Based on Extraction of Key Sentiment Sentence[M]// Natural Language Processing and Chinese Computing. Springer-Verlag New York, Inc. 2015.
21. Zhou G, He T, Wu W, et al. Linking Heterogeneous Input Features with Pivots for Domain Adaptation[C]// International Conference on Artificial Intelligence. AAAI Press, 2015.
22. 李鼎宇, 胡学钢. 面向短文本的跨领域情感分类算法[J]. 小型微型计算机系统, 2018, 39(05): 1005-1009.