学校代号 10530　　学　号 201431091146

分 类 号 O241　　密　级 公 开

湘潭大學

硕 士 学 位 论 文

基于引文分析和深度学习的新兴技术识别算法研究

学 位 申 请 人　何 春 辉

指 导 教 师　程 戈 副教授

学 院 名 称　数学与计算科学学院

学 科 专 业　数 学

研 究 方 向　信息处理及应用软件

二〇一七年四月十日

基于引文分析和深度学习的新兴技术识别算法研究

学 位 申 请 人　何 春 辉

指 导 教 师　程 戈 副教授

学 院 名 称　数学与计算科学学院

学 科 专 业　数 学

研 究 方 向　信息处理及应用软件

学 位 申 请 级 别　理 学 硕 士

学 位 授 予 单 位　湘 潭 大 学

论 文 提 交 日 期　2017-4-10

Emerging Technology Identification Algorithm of Study Based on Citation Analysis and Deep Learning

Candidate Chunhui He

Supervisor Professor Ge Cheng

College Mathematics and Computational Science

Program Mathematics

Specialization Information Processing and Application Software

Degree Master of Science

University Xiangtan University

Date April 10, 2017

湘潭大学
学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：　　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学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湘潭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进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

涉密论文按学校规定处理。

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘 要

新兴技术对企业的发展具有深远的影响，它可以引导企业跟上技术发展的趋势。众所周知利用新兴技术发展趋势可以帮助企业做出重要的技术和业务决策。因此完成新兴技术的识别工作就显得尤为重要，而现有的模型大多数是针对某些特定的技术领域进行识别，这通常会带有一定的局限性，无法满足实际的要求。鉴于此，本研究重点如下：

首先，对美国1975-2009年的专利引用数据进行了深入的分析，提取了一些对新兴技术识别具有代表性的特征项。根据以上提取的所有特征对专利引用数据进行了索引和聚类的预处理，接下来应用本研究提出的新兴技术和非新兴技术类别自动标注算法，实验结果表明类别自动标注算法取得了较好的效果。然后结合深度信念网络和逻辑斯蒂克回归模型构建了本研究提出的基于深度学习的新兴技术识别算法。

其次，使用了美国专利商标局提供的专利引用数据来测试和评估基于深度学习的新兴技术识别算法。该算法通过自动选取专利引文数据中最佳的特征组合来进行新兴技术的识别。测试表明，该方法能够准确和稳定地识别新兴技术，通过与其它算法进行对比，发现新算法的准确率比其它算法要高1.05%。

关键词： 深度学习；技术识别；限制波尔兹曼机；专利引文；特征抽取

ABSTRACT

Emerging technologies have a deep impact on the development of enterprises, It guide the enterprise to keep up with technological develop trends. Knowledge of emerging technological develop trends can help enterprises make important technical and business decisions. Therefore, it is very important to forecast emerging technologies yet most of the existing models identified for certain specific technical fields usually have some limitations and can't meet the actual requirements. In view of this, this study focused on the following :

First, with an in-depth analysis of the Patent citation data between 1975 and 2009, and extract some representative features of emerging technology identification. According to the above, extracted all the features of the patent citation data for the indexing and clustering and then applied the proposed emerging technology and non-emerging technology category automatic labeling algorithm, The experimental results show that the automatic labeling algorithm achieves good results. Then, based on the depth belief network and the logistic regression model, this study constructed the emerging technology recognition algorithm based on deep learning.

Secondly, used the patent citation data provided by the USPTO to test and evaluate emerging technology identification algorithms based on deep learning. The algorithm identifies the emerging technology by automatically selecting the best combination of features in the patent citation data. The results show that the proposed method can accurately and stably identify emerging technologies. Compared with other algorithms, the accuracy of emerging technology recognition algorithm is found to be 1.05% higher.

Keywords: Deep Learning; Technology Identification; RBM; Patent Citation; Feature Extraction

目 录

摘 要 I

ABSTRACT II

第一章 绪 论 1

§1.1 研究背景 1

1.1.1 新兴技术的概念及其特征 1

1.1.2 研究意义 3

§1.2 国内外研究现状 3

1.2.1 新兴技术的分类 3

1.2.2 新兴技术的识别方法 4

1.2.3 深度学习在识别领域中的应用 8

§1.3 研究内容 10

§1.4 论文结构 11

第二章 算法的理论分析 12

§2.1 聚类和分类方法概述 12

2.1.1 聚类方法 13

2.1.2 分类方法 15

§2.2 深度学习概述 21

2.2.1 深度学习的网络结构及参数训练方式 21

2.2.2 深度学习的训练策略 22

2.2.3 深度学习算法的评估标准及参数学习 22

§2.3 受限玻尔兹曼机 24

§2.4 本章小结 25

第三章　新兴技术识别算法的设计及构建 26

§3.1　特征选取 26

3.1.1　特征及其概述 26

§3.2　数据的预处理 29

3.2.1　索引 29

3.2.2　聚类 29

3.2.3　类别标注 30

§3.3　基于深度学习的新兴技术识别算法 34

3.3.1　构造数据矩阵 35

3.3.2　重构算法的选取以及参数学习 36

3.3.3　深度学习模型的反馈微调 40

§3.4　本章小结 41

第四章　算法评估准则及实验对比分析 42

§4.1　算法评估准则 42

§4.2　数据集的获取 43

4.2.1　网上收集数据 43

4.2.2　利用网络爬虫获取数据 43

§4.3　实验结果及对比分析 45

4.3.1　深度学习算法参数的训练 45

4.3.2　基于深度学习的新兴技术识别算法实验结果 49

4.3.3　新兴技术识别算法与其它经典算法的简单对比 50

§4.4　本章小结 53

第五章　总结与展望 54

§5.1　全文总结 54

§5.2　未来展望 55

参考文献 56
致 谢 62
攻读硕士学位期间的主要研究成果 63

图目录

图 2.1 新兴技术识别模型流程图 ····· 12
图 2.2 密度可达(左)和密度相连性(右)示意图 ····· 13
图 2.3 SVM算法流程图 ····· 18
图 2.4 双隐藏层神经网络结构图 ····· 19
图 2.5 多层前馈型网络结构图 ····· 20

图 3.1 聚簇类标签标注算法流程图 ····· 31
图 3.2 深度学习系统体系结构图 ····· 35
图 3.3 用于深度学习模型的数据矩阵 ····· 36
图 3.4 重构算法在不同迭代次数下的重构误差对比图 ····· 37
图 3.5 隐藏层节点数量与重构误差之间的关联关系图 ····· 37
图 3.6 深度神经网络各层之间RBM调节详细过程图 ····· 38
图 3.7 深度学习模型的整体反馈微调流程图 ····· 40

图 4.1 RBM迭代1次的实验结果 ····· 46
图 4.2 RBM迭代5次的实验结果 ····· 46
图 4.3 RBM迭代多次的实验结果 ····· 46
图 4.4 深度学习模型迭代1次的实验结果 ····· 47
图 4.5 深度学习模型迭代多次的实验结果 ····· 48
图 4.6 深度学习系统的重构误差变化图 ····· 48
图 4.7 深度学习系统的交叉熵变化图 ····· 49
图 4.8 新兴技术识别算法与传统算法的实验结果对比图 ····· 52

第一章　绪　论

§1.1　研究背景

回望人类历史的长河，所有工业革命的兴起都是离不开某种新兴技术的出现，谁率先掌握了这些技术，他们就夺得了发展的先机[1]。许多发达国家的迅速崛起，都与新兴技术的创新应用有着密切的关系。因此，沿着新兴技术的发展方向，识别出有前景的新兴技术，已逐步成为各国最关注的焦点。在传统互联网技术与新兴的大数据技术时代背景下，人类第四次工业革命也出现了雏形，中国要想成为科技强国，走上一条和平崛起的发展道路就必须抓住机遇大力推广新兴技术的发展，来实现科技兴国的中国梦，这是时代发展的趋势。

§1.1.1　新兴技术的概念及其特征

（一）新兴技术的概念

“技术”这个词语最初是来源于“technique”,专指“制和作的系统知识和技艺”。因为新兴技术这个概念问世的时间相对来说很短暂，所以工业界和学术界都没有给出完整的定义。

在快速发展的21世纪，科技的革新速度已经进入到一个前所未有的时代。新兴技术的发展势不可挡，呈现的类型也层出不穷。然而这些新出现的技术在过去都被笼统地称为“新技术”，早期学者对“新兴”定义为“现有市场中的应用在经历革新或新市场正在发展或形成”；“技术”被定义为“建立在行业基础上并适应于某一特定产品或市场的一类技能”。他们认为新兴技术应具有极强的时效性，也应当是正在逐步为人们所了解，但尚未成熟的技术。然而，从不同的维度来看待问题，就会有不同的见解，国外学者和国内学者在对新兴技术的概念探讨上也存在一些不同之处。

1.国外的定义

国外有学者认为新兴技术指的是“一种基于科学理论，在未来有可能创造出一个新产业或直接改善一个老产业的技术”[2]。众多学者都一致认为新兴技术的

范围应该涵盖带有突破性原创的间断性技术领域，以及集成过去相互独立的多项技术而形成二次创新的新技术领域。

2.国内的定义

有相关学者对新兴技术的概念给出了新定义，以银路为代表的一类学者将新兴技术定义为新产生以及正在发展中、且对市场和经济结构产生很大影响的一类高新技术[3]。华宏鸣认为新兴技术应该是指“当前还没被投入商业使用，但在未来几年内会被投入商业使用的技术”[4]。

本文的新兴技术符合国外学者对它的定义，从技术分类号的角度考虑，结合美国专利分类号构建了一个基于专利引文关系的新兴技术识别模型。

（二）新兴技术的特征

通过新兴技术与普通技术的对比分析，研究者得出新兴技术不仅具有普通技术的一般特征，此外还具有非常明显的特征，即高度不确定性。技术不确定性涉及市场的不确定性和技术本身的不确定性两个方面。通过深入的分析，银路[5]等提出新兴技术具有高度不确定性和复杂性以及创造性毁灭的多重特点。新兴技术创造性毁灭的特征是指它既可以单独开创一个新产业，也可以推毁市场中现存的老产业，或者改变组织的经营模式和竞争手段。具体特征如下：

1.高度不确定性

技术自身所潜在的不确定性，主要体现为新兴技术作为科学前沿的探索性热点技术，其研发成败的结果无法准确估计。外部环境的不确定性也会影响新兴技术的发展，例如政府的政策支持和宏观调控、替代技术的发展状况、产业链发展情况等都会产生相应的影响。

2.高度复杂性

与传统技术相比，新兴技术的渗透性较高，当前的绝大多数新兴技术，都不会单一的在一个领域内发展，在特定产业中，新兴技术的跨度非常大。对于这种通过多领域交叉发展起来的技术，其复杂性是相当高的。

3.创造性毁灭

由于新兴技术的发明和应用，创造新行业，毁灭一个甚至多个传统行业的现象随处可见，三次工业革命的出现就是深深的历史明镜。比如IT 技术的产生和发展，时刻都在创造着新的、巨大的行业，同时也使得一些传统行业不得不改变营销模式甚至是转投其他行业。

§1.1.2 研究意义

世界经济论坛给出了全球排名前五的新兴技术，大致为:“无线供电”、“创新研究新信息附加价值”、“生物工程”、“材料的纳米级设计”以及“计算机仿真建模”[6]。

新兴技术是近几年才提出的新概念，它的研究始于20世纪90年代。通过对相关文献的分析，得出新兴技术拥有三个核心的特征，依次为:“技术知识的可扩展性”、“现有市场的革新性”、“新市场的形成性”[7] 。中国在国家科学技术发展规划中明确提出要大力推动新兴技术的发展与应用。在新兴技术识别中，目标技术和新兴技术的依赖性起到了关键的作用，并且技术发展越快，新兴技术的作用就越突出。正因如此，在所有新技术中对新兴技术进行有效识别就显得至关重要，它将直接关乎到中国的经济、科技的发展速度。

随着社会的全面发展，各大领域里的新兴技术快速的涌现出来。但是真正能够进入市场并产生较大社会影响的却是寥寥无几，因而，谁能率先识别并应用这些技术指导生产实践，谁就能在竞争中脱颖而出，从而引领群雄。在经济、信息全球化和国家大力提倡创新的大背景下，大力培育和发展新兴技术对推进中国现代化建设具有特殊的现实意义。

§1.2 国内外研究现状

§1.2.1 新兴技术的分类

随着技术的发展，以多学科融合为背景的技术革新给人类日常生活带来了巨大的变化。他们促使现有市场格局和竞争规则不断发生变化，与此同时也带来了大量的发展机遇，为人类创造了无数的新产业[8]。

研究人员按照“技术应用领域”和“技术融合范围”[9]以及“产业化时间段”和“行业影响力差异”[10]四个维度对新兴技术进行了相关划分，大致可以分为以下四类：(1)突变型新兴技术，这种技术是指在现有应用上取得新突破的一类新兴技术；(2)植入型新兴技术，此类新兴技术通过将新技术集成到原有的应用中去，从而形成新的技术创新，因此它通常包含两项或两项以上的技术融合；(3)应用创新型新兴技术，就是将一项新兴技术推广到其他的应用领域从而形成二次创新的新技术；(4)融合型新兴技术，指通过多项技术的相互交叉及融合，并在新的领域得到快速应用和发展的新兴技术。以上这四种分类方法是目前国际上研究人员采用较多的分类方式，除此之外，还有少部分学者对新兴技术的分类持有不同的看法，但是都未得到广泛的认可。

§1.2.2 新兴技术的识别方法

新兴技术识别是指在众多的新技术中，通过合理有效的手段发掘出对社会生产有重大指导意义的技术和模式。随着社会的发展，新兴技术识别的手段和方法越来越多，复杂性也越来越高，识别难度也在逐步增大。技术识别已经成为了一个综合选择与评估的过程，下面按照所使用的基本方法分别进行阐述。

1.主观识别方法

最早的新兴技术识别方法主要采用专家讨论的形式来实现，此方法虽然简单，但存在“蝴蝶效应”等不足之处[11]。德尔菲发挥了群体决策的优势，这样可以涵盖领域内大多数专家的意见，避免面对面的讨论，比之前采用的个体决策方式相比具有更好的效果。虽然德尔菲法提供了一个完整集中不同专家意见的平台，但它成本高、费时，且问题的模糊性和不确定性仍然存在[12] 。考虑到以上问题，之后提出了模糊德尔菲方法来改进它，以弥补上述的弊端[13]。

在模糊德尔菲法发展的同时，萨蒂[14]引入了层次分析法来识别新兴技术,它自引进以来，广泛被用于新兴技术识别领域，随着层次分析法与德尔菲法的快速发展，基于他们两者的组合方法－－德尔菲-层次分析法（DHP）的出现，为新兴技术的识别提供了一个新方向。技术的选择会深深的影响一个企业或国家的发展，因此通过集成不同方法来形成多准则的混合决策方法是一个重要的策略，这种方法相对于一般的主观方法来说具有很大的优势，它可以在一定程度上消除主

观性，帮助决策者做出更合理的选择。随着技术数量和复杂性的增加，正确识别新兴技术变得越来越难。Shen Y C [15]提出了一种集成模糊技术的选择方法，并通过对二极管技术的识别来验证了模型的有效性和可行性。

2.基于文献的识别方法

新兴技术识别面临的最大问题之一就是怎么从海量的新技术中识别出具有发展潜力的新兴技术。很多研究成果[16, 17, 18]已经证实，基于已有文献可以为研究人员提供全面了解一个特定领域的发展现状，并有效的帮助他们识别出具有市场前景的新兴技术。

①基于专利文献的识别方法

近年来，涌现了一大批基于专利文献特点来识别新兴技术的方法。通常从专利文献入手，根据相关理论建立框架，融合文本挖掘、主题聚类、网络演化等技术手段对新兴技术的识别进行研究。

其中比较具有代表性的有：陈亮[19]基于专利文本中术语共现关系，采用同质块建模方法对连续时间段术语共现网络的变化情况进行研究，以识别新兴技术的系统构成，并以硬盘驱动器领域磁头技术为例进行了实证分析且取得了很好的识别效果。李倩[20]基于专利信息抽取和模式识别理论从纷繁复杂的技术发展初期对弱信号进行了识别，并发现了潜在的技术发展趋势。王凌燕[21]在专利主题基础上提出了新兴技术识别的初步框架，并对生物领域新兴技术识别做了实证分析。

C. Eusebi和R. Silberglitt [22]以美国专利分类系统提供的数据为基础对专利之间的关系进行分析，提出利用专利分类体系开展技术预测的新方法。在不同的领域，新兴技术都起到了发展和创新的推动作用，然而实现新兴技术的识别与预测是一项艰巨的任务，因为目前大部分都是依靠主观的专家经验来判断，这会影响客观事实，为了避开这类问题的出现，Jun S[23]提出利用专利客观分析法来实现新兴技术的识别，该方法用过专利文献进行分析，结合数据挖掘技术从专利的标题和摘要中提取关键词并结合专利的IPC 分类号来分析预测新兴技术。碳纳米显示技术是显示领域的重大创新应用。因此预测该新兴技术的发展趋势就成为了业内一个重要的课题。

Chang P L [24] 提出利用专利文献计量分析和专利网络分析方法来模拟碳纳米显示技术的发展趋势。该方法首先是根据该领域现有的专利情况构建一个

网络，然后利用网络分析方法对构建的网络进行分析，得出新兴技术的发展趋势。Yoon B[25]认为形态学分析确实是一个不错的方法，但它还是有一些不足之处。为了弥补这些不足，他提出了一种结合形态学分析和联合分析以及专利分析来实现新兴技术识别的混合新方法。该方法首先通过文本挖掘技术来提取关键词构造新兴技术的预定义矩阵，然后使用联合分析的方法对预定义的新兴技术进行筛选，最后才利用专利信息来确定最终的新兴技术。经过实证，该方法在新兴技术的识别准确率上有较大的提升。

②基于非专利文献的识别方法

除了部分学者利用专利文献来识别新兴技术之外，也有一些研究人员通过非专利文献分析的技术手段来识别新兴技术。在“机械和材料”领域里处于前瞻性的20种新兴技术被应用在短期预测的研究上。对于一个给定的技术问题，Bengisu M[26]认为科学计量学可以发挥出最佳关键词链接的优势来确定这些领域的专利和出版物数量，而这些出版物与大部分的新兴技术研究是高度相关的，该方法提出了一种简单而高效的用于新兴技术研究经费管理和投资的决策方法。在没有历史数据的情况下，很难对新兴技术进行预测，在这种情况下，利用文献计量学提供的数据来研究是一个不错的选择，Daim T U[27]通过整合文献计量学的资源并利用情景规划、增长线、类比法等提出了一个新兴技术识别模型，然后在燃料、食品安全以及光学存储技术3个领域验证了模型的有效性。

③基于专利引文分析的识别方法

引文分析经过长期的发展，尤其在专利引文方面奠定了深厚的理论基础。技术竞争主要体现在专利上面，专利仅以期刊总资源十分之一的比例，囊括了全世界百分之三十以上的新产品信息[28, 29]。新兴技术识别是关系到一个国家和地区未来发展的重要战略议题，对于许多企业来说，能提前掌握新兴技术的信息，就可以大大的降低风险。在1999年，有学者提出了新兴技术和专利引用频次之间存在某种关系的观点。

Leydesdorff[30]根据专利引文数据、IPC分类号、德温特分类号构建了一种新兴技术相似度的测量方法，并以美国授权专利为样本进行实证分析，发现该方法对新兴技术相似性区分较好。Criscuolo[31]利用欧洲专利局和美国专利商标局授权的专利引文数据库为数据源进行实证分析，通过对专利引文信息进行研究，找

出了相似的新兴技术。

李睿[32]从专利引文的时效性和实际可操作性方面对这两种方法进行了对比分析，发现引用耦合方法在揭示新兴技术的相似性方面更具优势，同被引方法则更适用于发现基础技术的演化规律。Holman[33]依据新兴技术和专利文献的核心特征，建立了基于多级专利耦合聚类的识别算法及相关评估指标。专利引用的网络连接是一个进化图, 展现了一个创新的过程。érdi[34]提出了利用引用向量作为预测器来对新兴技术进行预测，利用欧式距离来计算专利相似性，对美国专利商标局（1975-2007）子类11下面的所有专利的引用信息进行了实证分析，通过系统的结构树形图来分析和预测新兴技术。Ta-Shun Cho[35]从国家战略角度出发提出有效识别新兴技术对制定相关政策有至关重要的作用。他通过对美国专利商标局授权的专利文献使用网络引文分析的方法识别出了台湾地区在1997-2008年之间的新兴技术。

3.可视化识别方法

在信息时代，许多新技术已经被应用到新兴技术识别研究中。刘倩楠[36]收集了某个特定领域的专利引文数据并应用识别算法和可视化方法体系，较好的预测到了新兴技术的可能发展方向。王贤文[37]采用科学计量学中的相关技术，通过构建大型的关联矩阵，识别出了新兴技术的网络结构。Erdi[14]认为一项专利引用另一项意味着引用专利承载了被引专利的部分知识。他基于USPTO分类体系构建了网络，从而来讨论新兴技术的产生、成长、收缩、分裂、重组以及消亡的演化过程。利用图论的相关理论并结合可视化方法对新兴技术的演化过程构造了树状图体系，通过树状图的细微变化来预测新兴技术。Kim[38]认为可视化方法在呈现新兴技术分析结果方面具有不可替代的优势，然而，目前的专利地图可视化方法在新兴技术的研究中还存在一些不足，为了克服专利地图的不足，他引入了语义网络的知识体系，从专利中提取关键词来构建新的网络节点，从而可以形成新的专利地图，新地图里包含了更多的特征信息，可以更清晰的帮助人们掌握新兴技术的变化趋势。

4.其它的识别方法

近些年来，利用专利来识别新兴技术的方法确实得到了很大的发展，但不可否认的是，这些方法的可扩展性较差，无法解决众多领域的识别任务。为了弥补

这个缺陷，有学者根据不同的需求和目的采用特定的方法来识别特定的新兴技术。

黄鲁成[39]依据属性测度原理提出了属性综合评估和决策系统，对技术识别给出了简单的判别模型，以识别其中的新兴技术。李欣[40]意识到专利引用信息存在滞后性且采用关键词聚类不能全面反映技术主题的缺陷，提出了基于SAO（subject-Action-Object）的新兴技术识别方法。王鹏[19] 从经济维度、科技维度和空间维度进行切入，提出了战略性新兴技术的识别模型，并对中国风电产业的战略性新兴技术进行了识别，证明了模型的有效性。专利文献是技术和商业最主要的知识资源，因此专利分析已经被认为是管理技术经济的最有用的方法之一。虽然基于专利的新兴技术研究方法占据了上风，但是Yoon B[41]认为专利引文分析方法还是存在一定的缺陷，为了避免不足，他深入分析了专利网络发展的全过程，提出了一些新的指标，例如技术中心指数、技术周期指数等等来衡量专利网络的可靠性。通过引入这些新的指标，就显示了一个全新专利关系网络，该网络可以更加直观和有效的分析和识别新兴技术的发展方向。

§1.2.3 深度学习在识别领域中的应用

深度学习是机器学习里面最有活力的新领域，通过提升硬件的计算性能，深度学习算法已经得到了普及，近几年它在语音和图像识别、行为识别等领域都有巨大的应用[42]。

1.深度学习在语音识别中的应用

随着技术的成熟，LiDeng[43]指出深度学习在工业界中已经成为了语音识别领域的主流技术，并对目前的深度学习技术的基本功能和局限性进行了系统的阐述，最后沿着传统方法的特征和模型路径对语音系统展开了分析，对相关技术未来的应用方向进行了探讨。陈硕[44]利用RBM堆叠构建深度信念网络模型，并用于非特定人的语音识别，在DBN中，通过对时间进行处理之后得到的MFCC一阶差分参数作为模型的输入，模型的学习效果得到了一定的提升，使识别准确率有了较大的提高。

2.深度学习在图像识别中的应用

有学者指出，深度学习可以通过一种多层非线性网络节点结构实现复杂函数的逼近。林妙真[45]依赖深度学习的方法提出了基于多姿态的脸部识别算法并在脸部识别上获得了较好的效果。由于深度学习的参数训练难度比较大，K He[46]，提出了一种基于残差的深度学习框架，通过该框架不仅可以大幅降低网络的训练难度，而且还可以达到加深网络结构的目标。他们通过结合图像识别领域的具体实例给出了实证分析，在业界公认的Image Net数据集上应用该方法来实现深度学习算法，实验效果较好，和一般的方法相对比，准确率高出了3.57%，也因此贡献，整个研究团队获得了2015年的分类任务ILSVRC最高荣誉的奖项。

3.深度学习在人类行为识别中的应用

为了实现对人类行为的识别，Baccouche M [47]提出了一个完全自动化的深度学习算法，在不使用任何先验知识的前提下，自动对人类的行为进行分类。首先是使用卷积神经网络来自动学习特征，然后使用递归神经网络为每一步得到的演化特征进行学习并给出分类结果，最后通过与现有方法的实验结果进行对比，给出了该方法的优点和最佳分类结果。在这之后，Xie L[48] 提出了一种基于视觉传感器的深度学习锥形人类行为识别架构。该方法包含三个步骤：首先是人物行为的特征预处理；接下来是创建一个深度隐藏神经网络；之后就是模式识别部分。作者利用堆栈式自动编码器来构建深度神经网络模型，然后通过网络全连接的方式来训练每个行为模式分类所需的参数最佳权重，最后结合人体行为进行相关识别实验，结果表明该方法效果较好。

通过对收集的文献进行归类分析，可以清晰的了解到国内外研究人员在新兴技术识别和深度学习及其应用中的研究成果。从最初的主观识别方法中，可以了解到它们有一定的缺陷，那就是人为因素在某些时刻会起到决定性的作用。为了克服这种缺陷，研究人员提出了基于文献内容的识别方法，它包括专利文献和非专利文献两类，通过从这些文献中抽取特征词来构成实体，然后在构建识别模型，这种方法在一定程度上降低了主观性的影响，但是特征词抽取的难度较大，而且会造成信息损失。为了降低特征抽取的难度，有学者提出了基于引文数据来识别新兴技术的观点，这种方法有很多优点，它既降低了主观性的影响，又简化了特征提取的复杂度，而且还有良好的可扩展性，这种方法成为了最近几年的研

究热点。就目前来说，深度学习的理论已经比较完善，并且在很多领域都得到了较好的应用，但在新兴技术识别领域还未曾有相关的研究成果，因此经过综合分析，本文最后选择利用专利引文分析的优势，并结合深度学习来构建新兴技术识别算法，这可以降低主观性的影响，还可以简化特征提取的难度。

§1.3 研究内容

通过总结发现，大多数研究成果表明，为了实现不同的识别任务，一般都会建立特定领域的识别方法，因此这些方法无法被应用到其它领域，这样便不利于新兴技术的发展。为了解决这个难题，本文尝试利用引文网络的优点来克服这种特定领域的局限性，并结合机器学习算法的优势来构建一个通用的新兴技术识别模型。模型的构建和验证过程大致分为以下几部分：

1）数据获取及数据预处理过程是验证模型有效性的直接影响因素，因此数据获取及预处理步骤是模型的核心部分；首先是从网上获取美国专利商标局自1975-2009年间已授权的发明专利的引文数据，接下来，通过对引文数据进行深层分析，选出具有代表性的特征项，然后再使用Java编程技术和Sqlite3数据库操作技术对获取的数据进行预处理，得到只包含选定特征项的特征矩阵文件，再利用聚类算法按不同年份对经过预处理的特征矩阵文件进行聚类，形成验证模型的训练和测试样本。

2）新兴技术识别算法的设计与实现；高性能计算机集群技术与深度神经网络学习算法为机器学习的发展提供了技术支持，通过对多种机器学习算法进行研究和对比之后，本文采用了结合深度信念网络和逻辑斯蒂克回归算法来构建新兴技术的识别算法，最后通过Python编程技术来实现基于深度信念网络的新兴技术识别算法。

3）算法性能的检验；本文采用美国专利商标局自1975年-2009年已授权的专利引文数据来验证算法的性能，然后通过对不同机器学习方法得到实验结果的准确率进行简单的对比分析，从而对算法性能进行评估。

§1.4 论文结构

本文的结构分成五个章节：

第一章 绪论：主要介绍了论文的选题背景和新兴技术的概念及其特征，然后简要的总结了国内外专利引文分析在新兴技术识别中的应用和新兴技术分类以及识别方法，并且概括了本文的主要研究内容和论文的主要框架结构。

第二章 算法的理论分析：主要介绍了聚类、分类学习方法和深度学习的基本理论以及参数的训练方法及策略，并且对受限玻尔兹曼机算法进行了简述。

第三章 新兴技术识别算法的设计及构建：主要介绍了特征选取以及数据的预处理过程，并且给出了基于深度学习的新兴技术识别算法的设计及构建原理。

第四章 算法评估准则及实验对比分析：主要对构建的模型进行了检验和评估，介绍了数据集的获取方式，然后给出了算法的实验结果，并将实验结果与其它机器学习算法的实验结果进行了简单对比。

第五章 总结与展望：对论文中得到的相关研究结果进行概述，并且剖析了其潜在的一些问题，最后简单介绍了未来的研究方向。

第二章　算法的理论分析

§2.1　聚类和分类方法概述

基于深度学习的新兴技术识别算法的理论部分主要体现在分类和聚类算法实现上面。其核心思想是：以年为基本单位，利用专利的主分类号和引文信息进行聚类，将相似技术划分到同一个聚族，这样就可以得到大量的重组聚族，接下来就是对这些重组聚族进行识别，如果把重组聚族的标签定义为新兴技术和非新兴技术的话，那么识别任务就可以看成是一个二类的分类问题。图2.1是通用的新兴技术识别模型流程图。

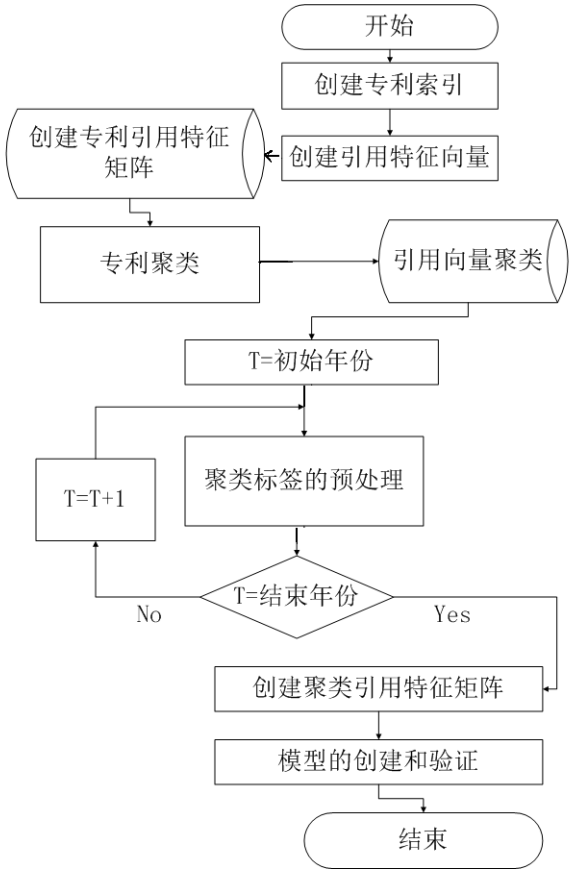

图 2.1　新兴技术识别模型流程图

§2.1.1 聚类方法

聚类是根据数据相关特征，将数据对象进行分组的一个过程。如果同簇内的对象相似性很高，异簇间的对象相异性很高，就说明聚类效果较好。

①DBSCAN聚类算法

DBSCAN是基于密度的通用聚类算法[49]。它不仅速度快且可以有效地应对噪声、无需预先给定聚簇数目，且对聚簇形状没有要求。它虽有诸多优点，但也存在两个明显的弊端：(1) 算法内存消耗与问题规模成非线性正相关关系；(2) 聚类结果的好坏与聚簇的概率密度分布均匀程度密切相关。图2.2 表示密度可达和密度相连性。

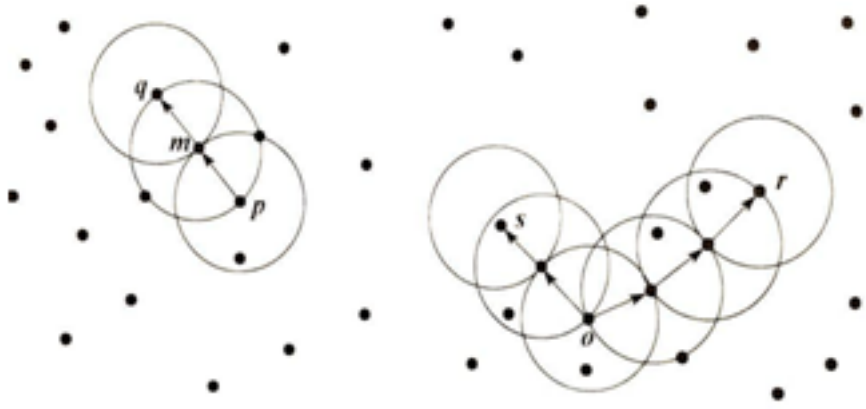

图 2.2 密度可达(左)和密度相连性(右)示意图

算法的概念及定义如下：

直接密度可达：假设p为选定的核心点，在以p为圆心，ε 为半径的邻域内所有点都是从p直接（密度）可达。

密度可达：对于点p，如果存在序列p_1, p_2,...,p_n,满足p_1=p，p_n=q，即p_i（$i \in [1, n)$）到p_n是直接（密度）可达的，那么p到q是（密度）可达的。

噪声：在簇外的对象就被划分为噪声。

算法步骤：

输入：样本数据D，初始化所有点为未访问，半径 ε 和最少点数MP。

S1：建立队列；

S2：如果D中数据全部处理完，则算法结束，否则，从D中选择一个未处理的点，标记为已访问，获得其所有直接密度可达点，如果为非核心点则标记为噪声，

重复步骤2，否则生成新的簇，进入步骤3；

S3：将当前核心点放入该簇，并将核心点的直接密度可达点放入队列，并遍历该队列，如果队列全部遍历完则回溯至步骤2；

S4：如果该点已经访问过，则进入步骤5，否则标记为已访问，然后获得该点的所有密度可达点，如果这个点也为核心点，则将该点的所有直接密度可达点放入队列；

S5：如果该点不属于任何簇，则放入当前簇；

S6：输出聚类结果,算法结束。

②K-Means聚类算法

K-means是一种常用的基于划分的聚类方法[50]。K均值算法初始化时要指定聚簇个数K。聚簇的相似度是对簇中所有对象相似度进行均值化，可以看作簇邻域的中心点。K-means 算法也表现出几个缺点：(1) 正常情况下，算法循环次数，要少于数据的总量。有时候会出现最坏的情形，使算法的时间复杂度比期望值要高。(2) K值的选取。在执行程序前，需要给定K 值的大小。然而对于不同的K值，划分的结果不同，因此确定合适的K非常关键。(3) 质心初始化。质心的初始选取对于划分结果也非常关键。

K均值算法流程大致如下：初始化时会随机的选取K个对象作为聚簇中心点，每个对象代表一个簇的初始均值或初始中心，对剩余的所有对象，计算它到各个簇中心的距离，并根据距离将它划分到离它最近的簇；然后计算每个新簇的均值。这个过程不断重复，直到准则函数收敛。通常，采用平方误差准则，其定义如下：

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{2-1}$$

式2-1中，E是所有距离的误差项总和，p是数据中的点，m_i是簇C_i中心的值。这个准则可以保证得到的K个簇内的对象是尽可能的相似。

K均值算法的步骤：

算法：K均值。用于划分K均值算法，每个簇的中心用对象的均值表示。

输入：

K: 簇的数目，

D：包含n个对象的数据集。

输出：K个簇的集合。

方法：

(1) 从D中任意选择K个对象作为初始簇中心；

(2) 循环内容：

(3)　根据簇中对象的均值，将每个对象指派到最相似的簇；

(4)　更新簇均值，即计算每个簇中对象的均值；

(5) 终止条件：聚族指派不再发生变化

§2.1.2　分类方法

分类是指通过对相关数据的特征进行自动学习，然后得到一个分类器F，最后把属性集X映射到一个预先定义的类标签Y的过程，分类器F被认为是一个分类模型。

①逻辑斯蒂克回归算法

Logistic Regression（逻辑回归）是机器学习中的一个常见模型，因为它具有简单高效的优点，在实际生产中也常常被使用，是一种经典的分类模型[51]。对于逻辑回归来说，其思想是在广义线性回归模型的基础上发展而来，公式如下：

$$h_\theta(x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\theta^T x}} \tag{2-2}$$

其中，$y = \frac{1}{(1+e^{-x})}$称作Sigmoid函数，算法本质上是将线性函数转化成Sigmoid函数。该函数的输出值介于(0,1)之间，即$h_\theta(x)$的输出是介于(0,1)之间，也就表明了数据属于某一类的概率，例如：$h_\theta(x) \leq 0.5$,则说明数据属于A 类，$h_\theta(x) > 0.5$则说明数据属于B类，因此可以将Sigmoid函数看成是样本数据的近似概率密度函数。根据公式(2-2)可知，θ 的值有着特殊的含义，它表示$h_\theta(x)$结果取1的概率，因此对于输入x的分类结果为1和0的概率分别为：

$$P(y=1|x;\theta) = h_\theta(x),$$
$$P(y=0|x;\theta) = 1 - h_\theta(x).$$

根据上式的结构使用极大似然估计来求解损失函数，首先得到概率函数为:

$$P(y|x;\theta) = (h_\theta(x))^y * (1-h_\theta(x))^{1-y} \tag{2-3}$$

因为样本数据独立，所以它们的联合分布可以表示为各边缘分布的乘积，取似然函数为：

$$L(\theta)=\prod_{i=1}^{m}\left(h_\theta(x^{(i)})\right)^{y^{(i)}}*\left(1-h_\theta(x^{(i)})\right)^{1-y^{(i)}} \tag{2-4}$$

取对数似然函数：

$$l(\theta)=\log L(\theta)=\sum_{i=1}^{m}\left(\log\left(\left(h_\theta(x^{(i)})\right)^{y^{(i)}}\right)+\log\left(\left(1-h_\theta(x^{(i)})\right)^{1-y^{(i)}}\right)\right) \tag{2-5}$$

最大似然估计目标是求出使$l(\theta)$取最大值时对应的θ，一般考虑选择梯度上升法进行求解。

②贝叶斯算法

贝叶斯定理[52]是250多年前的研究成果，在信息领域有重要的地位。贝叶斯分类是以条件概率为前提来进行分类的一系列算法的总称。朴素贝叶斯算法是最常用的分类算法之一。优点是：1、算法起源于统计和数学学科，有深厚的理论依据，以及稳定的效果。2、模型需要估计的参数较少，对缺失数据不敏感，算法简单，求解速度快。此外由于它有着很强的条件假设，所以它也会存在一些不足之处：1、理论上，NB模型与其他分类模型相比具有最小的误差率。实际上存在偏差，因为算法的前提条件是假设特征之间相互独立，这会给模型的分类结果带来不少的影响；2、需要知道先验概率。

贝叶斯定理的公式如（2-6）所示：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{2-6}$$

式中：P(A)是A的先验概率，$P(A|B)$被称作A的后验概率，$P(B|A)$被称作B的后验概率，P(B)是B的先验概率，也作标准化常量。

证明：根据条件概率的定义。在事件B发生的条件下事件A发生的概率是：

$$P(A|B)=\frac{P(A\bigcap B)}{P(B)} \tag{2-7}$$

同样地，在事件A发生的条件下事件B发生的概率是：

$$P(B|A)=\frac{P(A\bigcap B)}{P(A)} \tag{2-8}$$

整理与合并这两个方程式，可以得到：

$$P(A|B)P(B)=P(A\bigcap B)=P(B|A)P(A) \tag{2-9}$$

这个引理有时称作概率乘法规则。上式两边同时除以非零概率P(B)，可以得到贝叶斯定理的结论:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$ （2-6）成立，证毕。

算法原理及过程：

1.算法原理

设每个数据样本用一个n维特征向量来描述，即：X=(x_1,x_2,...,x_n),假定有m类，分别用C_1, C_2,...,C_m表示。给定一个未知的数据样本X（即没有类标号），若分类算法将未知的样本X 分配给类C_i，则一定满足条件：

$$P(C_i|X) > P(C_j|X)\ , \quad 1 \le j \le m, j \ne i$$

2.算法过程：

根据贝叶斯定理

由于P(X)对于所有类为常数，最大化后验概率$P(C_i|X)$可转化为最大化先验概率$P(X|C_i)P(C_i)$。如果训练数据集有许多属性和元组，计算$P(X|C_i)$的开销可能非常大，为此，通常假设各属性的取值互相独立，这样

先验概率$P(x_1|C_i), P(x_2|C_i), ..., P(x_n|C_i)$可以从训练数据集求得。

根据此方法，对一个未知类别的样本X，可以先分别计算出X属于每一个类别C_i的概率$P(X|C_i)P(C_i)$，然后选择其中概率最大的类别作为其类别。

算法成立的前提是各属性之间互相独立。当数据集满足这种独立性假设时,分类的准确度较高，否则可能较低。另外，该算法没有分类规则输出。

③支持向量机

支持向量机[53]通过映射，将原来低维数据映射到较高维的空间上，在新的空间上，它利用支持向量和边缘发现最佳分离超平面所在的位置，从而达到对数据集进行分类的目标。SVM的优点：可以提高泛化性能；可以解决高维问题；可以解决非线性问题（需构造核函数）等待。SVM的缺点：对缺失数据敏感；对非线性问题没有通用的解决方案，需要根据经验来选择相关核函数进行处理。

支持向量机的分类函数是一个简单的符号函数，$y = sign(w * x + b^*)$,$sign$,它的值由输入的符号来决定，根据法向量和截距以及输入特征向量，经运算后即可得到分类结果。为了能得到唯一的分界线，支持向量机提出间隔最大化的思想来确定超平面分界线，根据点到平面的距离公式分子为$|w * x + b|$,由于分母是相同的，所以$|w * x + b|$可以相对的表示距离的大小。y（向量）表示的是每个向量对应的类别，分类决策标准是根据函数来决定，可以确定$w * x + b$与y的符号（相对应元素）是一一对应的，由

$$y*(w*x_i+b) \tag{2-10}$$

来表示分类的正确性和确信度，$\overrightarrow{\gamma_i}$ 就是函数间隔:

$$\overrightarrow{\gamma_i} = y_i*(w*x_i+b) \tag{2-11}$$

$$\overrightarrow{\gamma} = \min \overrightarrow{\gamma_i}\ ,(i=1,...,N) \tag{2-12}$$

图2.3是SVM算法流程：

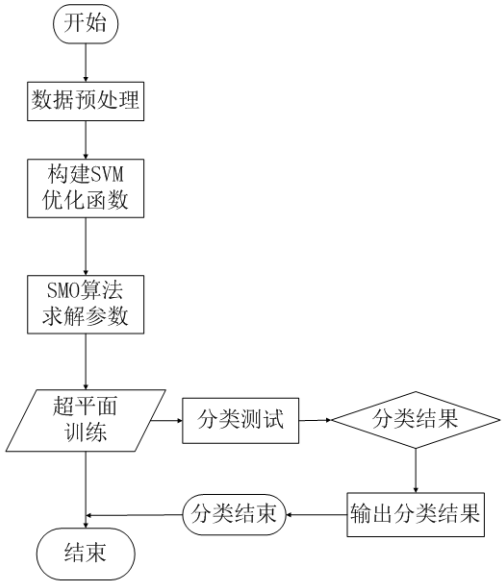

图 2.3 SVM算法流程图

④神经网络及其学习算法

1.多层感知机网络

多层感知机网络[54],在中间隐藏层节点上利用非线性函数进行逼近，得出数据从输入端到输出端的映射。如图2.4所示是一个拥有双隐藏层的神经网络。

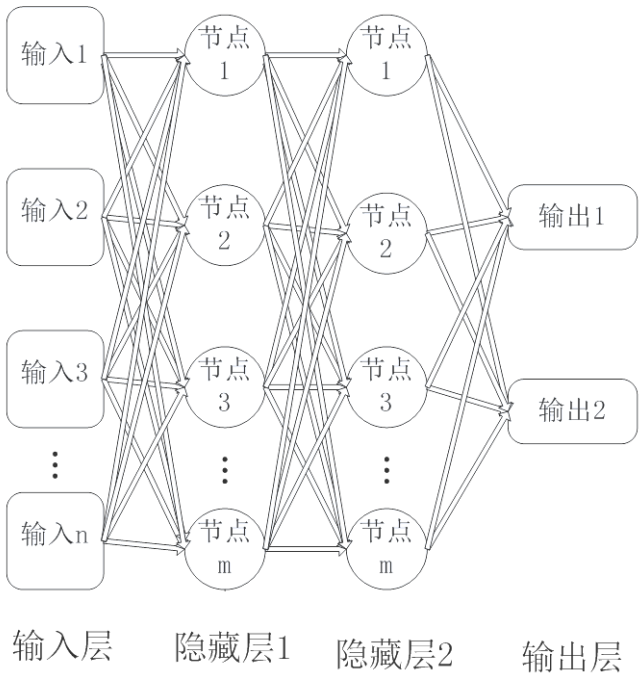

图 2.4 双隐藏层神经网络结构图

在图2.4中，网络由输入层、隐藏层及输出层构成，它能够表示种类繁多的非线性曲线，网络中使用的单元层越多，可以创造出的特征越复杂。输入层的输入向量$X = (x_1, x_2, \cdots, x_n)^T$，隐藏层输出向量为$Y = (y_1, y_2, \ldots, y_m)^T$，输出层输出向量为$O = (o_1, o_2, \cdots, o_l)^T$,期望输出向量为$d = (d_1, d_2, \cdots, d_l)^T$。输入层到隐藏层之间的权值矩阵用W表示，$W = (w_1, w_2, \cdots, w_j, \cdots, w_m)^T$；分向量$w_j$为隐藏层第j个节点所对应的权向量；隐藏层到输出层之间的权值矩阵用V表示，$V = (v_1, v_2, \cdots, v_k, \cdots, v_l)^T$；向量$v_k$对应隐藏层第k个神经节点的向量。

对于输出层，有：

$$o_k = f(net_k)(k = 1, 2, \cdots, l) \tag{2-13}$$

$$net_k = \sum_{j=0}^{m} v_{jk} * y_j(k = 1, 2, \cdots, l) \tag{2-14}$$

对于隐藏层，有

$$y_j = f(net_j)(j = 1, 2, \cdots, m) \tag{2-15}$$

$$net_j = \sum_{i=0}^{m} w_{ij} * x_i(j = 1, 2, \cdots, m) \tag{2-16}$$

激活函数f(x)都是采用单极性的Sigmoid函数，$f(x) = \frac{1}{(1+e^{-x})}$，该函数具有连续可导的性质，由以上各式就可以构成神经网络的数学模型。

2.BP学习算法

反向传播(简称BP网络)是另一种经典的神经网络算法[55]。一般情况下，神经网络是一组相互连接的输入/输出神经单元，其中每个连接都有一个权重。在学习阶段，通过调整这些权重，使得它能够正确预测出测试样本的类别。神经网络存在一些缺点：(1)需要很长的训练时间；(2)它需要训练大量的参数，这些参数只能靠经验来决定，如根据网络拓扑或“结构”来决策；(3)可解释性差。神经网络的优点：(1)对噪声数据的处理能力很强；(2)对未经训练数据集的分类能力较好。图2.5常见的前馈神经网络：

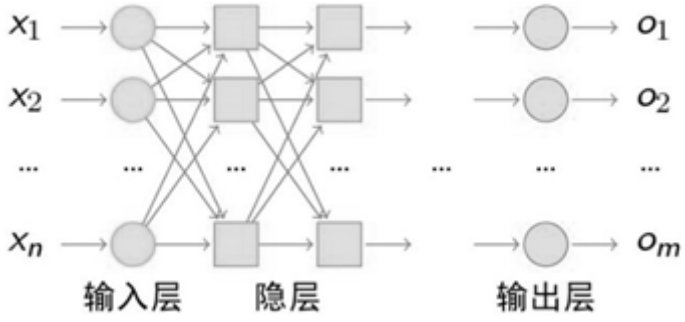

图 2.5 多层前馈型网络结构图

反向传播算法在多层前馈神经网络上学习。它通过循环迭代地方式来学习用于元组类别预测的权重。多层前馈神经网络一般是由三层构成。每层都有一些单元组成。网络是前馈型和全连接的，其参数值都不返回到输入节点或上一层的输出节点。每个单元都向下一层的单元提供输入。网络的输入对应每个训练元组的观测性质。输入同时提供给构成输入层的单元。这些输入通过输入层加权处理后直接提供给隐藏层使用，隐藏层单元的输出可以作为下一个隐藏层的输入，以此类推。隐藏层的数量是任意的。输出层发布给定元组的网络预测。

反向传播算法步骤：

（1）初始化：给权值和阈值赋予(-1, 1)之间的随机值；

（2）随机取一对样本对网络进行训练；

（3）计算中间层的输入/出；

（4）计算输出层的输入/出；

（5）计算输出层的一般误差；

（6）计算中间层的一般误差；

（7）修改输出层的权值和阈值；

（8）修正隐藏层的权值和阈值；

（9）取下一对样本，返回（3）开始训练，直到m个样本训练结束；

（10）判断全局误差是否小于预定值，否则，回到（2）重新进行训练，直到满足要求或达到预定训练次数，停止训练。

§2.2 深度学习概述

有一个系统S，它有n层S_1,...,S_n，它的输入是I，输出是O。假设系统通过参数的自动学习，使得输出O等于输入I，那么系统就可以获得输入I的一系列层次特征，S_1,...,S_n。

深度学习的理论思想是通过设置多个层次的网络，让机器自动提取特征，从而实现对输入信息进行分级表达。另外，前面是假设输出严格地等于输入，这个前提条件太严格，为了能更好的解决实际问题，可以略微地修改这个条件，很多时候只要使得输入与输出的误差极小就可以得到相应的模型。

§2.2.1 深度学习的网络结构及参数训练方式

对于深度学习，其使用的网络是一种多层前馈型神经网络。神经网络所对应的表达式通常是重组了多个函数而成的复合函数。如

$$f(x) = g_3(g_2(g_1(x; w_1); w_2); w_3) \tag{2-17}$$

函数f(x)由三个函数嵌套而成,三个函数各有自己的参数。

一个由嵌套函数构成的函数是非常复杂的，参数空间较大，很难找到全局最优解，从而使得这种模型效果还不如其他的简单模型。近年，学者终于找到了一种能够有效训练参数的方法，即：逐层预训练并进行最后的循环微调。

该方法需要构造一个自动编码器神经网络。经过自动编码训练后，记录它的输入层到隐藏层的连接权重，这些权重会作为深度神经网络输入层到第一个隐藏层权重的初始值。

根据以上方式，深度神经网络输入层到第一层隐藏层的权重就可以完成预训练。此时，构造另一个自动编码器神经网络，输入层是前面得到的深度神经网络

的第一个隐藏层，隐藏层是深度神经网络的第二个隐藏层。仍然按照刚才的方式训练自动编码器神经网络，让第二个隐藏层的信号能够尽量还原成第一个隐藏层的信号。如此训练自动编码器后，得到它的输入层到第一个隐藏层的权重，第一个隐藏层到第二个隐藏层的权重。依次逐层重复上面的步骤，直到所有层之间的权重均被初始化。最后用这些初始化后的权重，按照前一节的方法训练整个深度神经网络，这最后一步就称为微调。经过预训练后得到的神经网络，会明显优于随机给出初始权重训练得到的神经网络。

§2.2.2 深度学习的训练策略

传统的BP算法对于多层网络的训练效果不是很好[56]。BP算法的缺点：（1）梯度具有稀疏性：从上往下，误差校正信号逐级减少；（2）算法很难收敛到全局最优解；（3）只能对带有标注结果的数据进行训练：但大部分数据无类标签。

因此，有学者提出在非监督数据上建立多层神经网络的有效方法，该方法分为两步，一是每次训练一层网络，二是调优，使原始表示x向上生成的高级表示r和高级表示r向下生成的x'尽可能一致。方法是：

1）逐层构建网络节点，这样一次只训练一层网络所有节点的权值。

2）当所有节点的权值训练结束以后，用wake-sleep算法进行微调。

上形权重表示“认知”，下形权重表示“生成”，然后使用Wake-Sleep算法微调所有节点的权值。这样会让生成的顶层结果能够较好的复现底层的信息。W-S算法分为两个阶段。

1）认知阶段：通过外部特征和认知权重来产生抽象表示，并使用梯度下降法来逐步的优化下行权重。

2）生成阶段：通过网络的顶层状态和生成权值，生成底层的状态，同时逐步优化层间的向上权值。

§2.2.3 深度学习算法的评估标准及参数学习

①算法评估标准

采用机器学习领域的传统方法来评价深度学习模型性能是比较常见的做法。对于给定的一个输入，希望通过网络学习后能够得到期望的输出。设输入为X，

输出为Y，期望的输出为Y^*，可定义实际输出与期望输出的差别作为评价神经网络优劣的指标。一般情况可以将两个向量的距离的平方作为这个差别，即：

$$d(y, y^*) = \frac{1}{2} * |y^* - y|^2 \quad (2\text{-}18)$$

只考察单个样本还不够充分，通常需要考察一个样本集合，例如有：$(x_i, y_i^*)|i = 1, ..., N$,定义损失函数为：

$$L(w) = \sum_{i=1}^{N} d(y_i, y_i^*) \quad (2\text{-}19)$$

根据上述转化，最后采用$L(w)$来评估模型的优劣。

②参数学习

通过分析，发现损失函数可以用来评价模型的优劣，那么让损失函数的值达到最小的那组参数得到的值就应该是参数的最佳权值：$w^* = \arg min_w L(w)$,在数学上，这是一个无约束条件的优化问题，通过调整一组变量使得某个表达式的函数值最小（或最大），而对调整变量的取值是没有限制的。

为了解决上述无约束优化问题，梯度下降法通常是解决这一类问题的标准方法，它也是普遍用于神经网络参数权值优化的方法之一。梯度下降法是一种迭代方法，最初是任意选取一组参数，然后逐次的对这组参数进行微小的调整，使损失函数的取值逐步降低。

这个方法可以这样形象的理解，不妨设需要优化的参数只有两个，则参数空间是一个二维平面。任意一组参数对应于损失函数值，这构成第三维。这个三维空间形成一个曲面，如同高低不平的地形图，经纬度表示参数，高度表示损失函数的值。那么优化问题就是找到高度最低的经纬度。梯度下降法的思路是，首先任意选择一个地点，然后在当前点找到坡度最陡的方向，沿着该方向迈一小步，到达新的起点，开始进行下一轮迭代。通过不断的进行迭代，就会走到一个海拔最低的地方。所谓坡度在数学上就是指微分中的梯度，梯度下降算法描述如下：

梯度下降法：

1）初始化参数w_0,t=0

2）步数t=t+1

3）计算梯度$\nabla W = \frac{\partial loss}{\partial W}|w_{t-1}$

4）更新参数$W_t = W_{t-1} - \gamma \nabla W$

5）如果收敛，结束并输出W_t,否则转到步骤2。

§2.3 受限玻尔兹曼机

受限玻尔兹曼机[44](RBM)可被视为一个特殊的无向图模型，它是构成深度信念网络的基本模型。其中H为隐藏层，W为链接矩阵，V 为可见层。一般情况下都会假设所有的V和H均为二值变量，即$\forall i,j$,有$V_i \in_{\{0,1\}}, H_j \in_{\{0,1\}}$。

假设RBM网络总共有n个可见节点、m个隐藏节点，用向量v表示可见节点所处的状态，向量h表示隐含节点所处的状态。对于某组给定的状态（v,h），RBM堆叠系统所拥有的能量可以记为：

$$E(v,h|\theta) = -\sum_{i=1}^{n} a_{iv_i} - \sum_{j=1}^{m} b_j * h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i * W_{ij} * h_j \qquad (2\text{-}20)$$

上式中，$\theta = a_i, b_j, W_{ij}$是RBM的参数，他们均为实数。其中，$a_i$表示可见节点i的偏置，$b_j$表示可见节点j的偏置，$W_{ij}$表示可见节点i与隐藏节点j之间的链接矩阵，当参数确定时，基于该能量函数，可以得到（v,h）的联合概率分布，$P(v,h|\theta)=\frac{e^{(-E(v,h|\theta))}}{Z(\theta)}$, $Z(\theta) = -\sum_{v,h} e^{-E(v,h|\theta)}$, 其中$Z(\theta)$ 为归一化因子。

为了求解出RBM的联合概率中的边缘分布，$P(v|\theta)$，也称为似然函数：

$$P(v|\theta) = \frac{1}{z(\theta)} * \sum_h e^{(-E(v,h|\theta))} \qquad (2\text{-}21)$$

上式需要很大的计算量才能得到归一化因子$Z(\theta)$，然后才能得到$P(v|\theta)$的边缘分布。假设a_i, b_j, W_{ij}已经求出，想要求出求得，想要直接求出上述公式中$P(v|\theta)$依然难以实现。

根据RBM的特殊结构的设计（层内节点无连接）可知，当给定某层节点的状态时，与它相邻一层节点之间的状态条件分布是相互独立的，即：

$$P(v|h) = \prod_{i=1}^{n} P(v_i|h), P(h|v) = \prod_{j=1}^{m} P(h_j|v) \qquad (2\text{-}22)$$

意味着给定某个可见节点的状态时，此时第j个隐藏层节点的激活概率为：

$$P(h_{j=1}|\theta) = \sigma(b_j + \sum_j v_i * w_{ij}) \qquad (2\text{-}23)$$

其中$\sigma(\cdot)$为sigmoid激活函数，定义为：

$$\sigma(x) = \frac{1}{1+e^{-x}} \qquad (2\text{-}24)$$

求得所有隐藏层节点后，基于RBM的对称结构，可见节点的激活概率为：

$$P(v_i = 1|h,\theta) = \sigma(a_i + \sum_j W_{ij} * h_j) \qquad (2\text{-}25)$$

在完成上述一系列转化之后，可求出$P(v|\theta)$联合概率分布的边缘分布函数。

§2.4 本章小结

本章主要对本文所涉及的相关理论和算法进行了概述，包括新兴技术识别模型的构建流程图、聚类算法和分类算法、深度学习理论概述以及RBM和多层感知机神经网络的简述。首先，本章阐述了DBSCAN和K-means聚类算法，以及逻辑斯蒂克回归、朴素贝叶斯和支持向量机、多层感知机网络等分类算法的原理与流程；然后，对深度学习的基本思想和训练方式以及算法的评估方法进行了简单描述；最后，对用于构建深度神经网络的受限玻尔兹曼机算法进行了介绍。

第三章　新兴技术识别算法的设计及构建

§3.1　特征选取

在数据预处理阶段，从专利引文数据中提取用来刻画新兴技术和非新兴技术的特征是最重要的过程。这是本文研究的一个重点，因为，模型预测的准确率是极大的依赖于所选取的这些特征，如果选取的特征具有足够的代表性，那么模型的精度就会有较大的提升。经过深层分析，对于每一篇专利的引文信息，提取特征如下：如表3.1所示：

表 3.1　特征中、英文名称

特征英文名称	特征中文名称
Number of Claims	权利要求项数
Number of citations	引文总数量
Number of Citations with non-patent literature	非专利文献引文数量
Patent Class	专利分类号
Technology Cycle Time	技术生命周期
Cited Technology Similarity Index	被引技术的相似性指数
Cited patents Assignee Similarity Index	被引技术所有者平均相似性指数

§3.1.1　特征及其概述

①权利要求项数（NC）

专利通常保护的是一项技术方案或方法，与普通对象不同，其受保护的范围比较模糊，为了清晰界定专利权的涵盖范围，现代专利制度中明确要求给出专利权利要求的具体内容。在权利要求里面，发明人需要明确指出该专利所保护的权利范围，并分别使用权利要求项来详细展开。经过分析发现，核心专利的权利要求项数比较多，普通专利相对较低，因此选取权利要求的项数作为一个特征。

②引文总数量（Nc）

专利引文数量与专利本身密切相关，考虑到专利引用存在发明人自引和审查员引用两种。通过分析专利引文信息可以发掘潜在的知识流动方向和核心专利技术，也可以揭示技术发展路径以及专利所有者的关联信息。因此可以根据专利引文的信息，统计出每一篇专利的引文总数作为一项特征。

③非专利文献引文数量（NCNPL）

许多新技术的发明都是来源于科学研究的发现，而科学研究成果大部分都是以科技文献的形式来呈现，这就说明新技术的产生和非专利科技文献之间存在某种关系，它可以使用科技指数来反映技术的原创性。因此在所有专利引文中，出现非专利文献引文的数量就可以作为技术识别的一项特征。

④专利分类号(PC)

由于专利申请数量的增加，为了便于管理，规定每一篇专利在申请后，无论是否会被授权，它都会被分配一个或多个特定的分类号，而这个分类号一般情况下是可以直接反映出该项技术所属的技术领域。来自不同领域的技术，特征一般会有较大的差距，因此考虑专利类别这个特征来区分不同领域之间的差异。

⑤技术生命周期(TCT)

技术生命周期这是一个关于被引专利技术的平均值指标。在不同领域里的技术生命周期是不一样的，该指标可以反映出不同技术之间的更新快慢。例如互联网领域的技术一般情况下更新速度都比较快，即说明它的生命周期比较短。相对而言，研究人员得出结论：生命周期较短的技术领域比生命周期较长的技术领域出现技术融合从而产生新兴技术的可能性会更大。因此选取技术生命周期作为一个具有代表性的特征，它的计算公式如下：

$$\{TCT_i = median_j\{|T_i - T_j|\} \quad (3\text{-}1)$$

其中T_i是第i篇专利的申请日期，T_j是第i篇专利引用的第j篇专利的申请日期。

⑥被引技术的相似性指数(CTSI)

前面介绍过为了便于专利的管理，会按照不同领域进行分类，这样就可以根据专利的分类号来辨识引用技术和被引技术是否属于相似技术。而有研究人员做过相关研究，研究结果表明，来源于不同领域的技术重组更有可能生成新兴技术，换言之，以相似技术作为参考，创新出来的技术能成为新兴技术的概率较低。

美国专利分类系统对不同领域的技术进行了划分，一共有450 个大类和5600个子类。大类只是限定了大概的领域，而小类才会给出更具体的领域，在实际中往往是采用大类和小类相结合来共同构成专利的分类号。为了能够对被引技术相似性指数进行定量的分析，下面给出两个主分类号之间相似性计算的具体公式：

$$CS_{ij}=\begin{cases}0 & \text{，如果i和j大类和小类都不相同}\\ 0.5 & \text{，如果i和j大类相同,小类不同}\\ 1 & \text{，如果i和j大类和小类都相同}\end{cases} \quad (3-2)$$

然后，在美国专利分类号中，一项专利往往可以拥有几项分类号，因此需要求出两项专利分类号之间的平均相似度PCS_{pq}，以下是PCS_{pq}的表达式：

$$PCS_{pq}=\frac{(\sum_{i=1}^{N_p}\sum_{j=1}^{N_q}CS_{ij})}{(N_p*N_q)} \quad (3\text{-}3)$$

这里N_p和N_q表示专利p和专利q各自所拥有的分类号的数量。

最后，再来计算第x篇专利的被引技术相似性指数，指标的计算公式如下：

$$CTSI(x)=\frac{(\sum_{n=1}^{N}PCS_{xn})}{N} \quad (3\text{-}4)$$

此处，N是x所引用的专利总数，n是被x所引用的第n项专利。

⑦被引技术所有者平均相似性指数(CASI)

研究表明，当一个公司获得一项极具竞争力的核心发明时，为了避开它的竞争对手对该项核心发明造成威胁，这个公司会主动申请与该项核心发明相似的专利，这个现象在专利引文中通常是以专利权人的自引来衡量的。一项美国专利通常情况下有一个或多个专利权人，采取下面的公式计算两项技术的专利权人相似性指标：

$$AS_{pq}=\frac{(\sum_{i=1}^{N_p}\sum_{j=1}^{N_q}A_i*A_j)}{(N_p*N_q)} \quad (3\text{-}5)$$

这里N_p和N_q是专利p和专利q各自的专利权人的数量，并且有以下定义：

$$A_i*A_j=\begin{cases}1 & \text{，如果专利i和专利j的专利权人是相同的}\\ 0 & \text{，否则}\end{cases} \quad (3-6)$$

考虑到有些专利的专利权人数量较少，若按照上面的公式来计算对他们来说就显得不是很公平，为了降低这种影响，提出了被引技术所有者平均相似性指数(CASI)，对专利x以及它的引用专利而言，使用AS_{pq}来计算CASI，定义如下：

$$CASI(x)=\frac{(\sum_{n=1}^{N}A*S_{xn})}{N} \quad (3\text{-}7)$$

这里N是专利x的所有引文总数量，n是专利x引用的第n篇专利。

§3.2 数据的预处理

经过整合网络下载和网络爬虫所获得的原始专利引文数据，可以获得美国专利商标局在1975-2009年之间已授权专利完整的引文数据，并存储在sqlite3数据库中。然后，利用数据库的SQL查询语句并结合之前选出的特征项，对原始引文数据表进行查询操作，构建出列属性中只包含七项特征的引文数据表。

§3.2.1 索引

由于数据库中表格的记录数太大，直接操作数据库的时间复杂度会相当高，为了能够高效的处理这些引文数据，使用开源的索引工具Lucene 来处理专利引文数据项的索引任务。它将引文数据表格的七个列属性看成一篇文档中的七个不同的域进行索引，然后通过对索引文件进行处理，这样可以大大的提高数据预处理的效率。

§3.2.2 聚类

为了更好的解决模型中引用专利相似度的计算问题，本文对只包含特征项的引文记录数据进行了聚类处理，这样就可以把相似的技术整合到同一个聚簇中去，最后通过检验这些聚簇的类标签就可以得出相关的结论。

前面对两种很常见的聚类方法给出了简单的介绍，即：DBSCAN和K-means。它们有自己的特色，前者是一种基于密度的聚类算法，在聚类之前不需要指定聚簇的数目K，在完成聚类之后，算法会自动给出聚类的结果和聚簇数目；后者是一种基于划分的聚类算法，它要求在聚类开始前给定聚簇的数量K，如果K值不合适，那么聚类结果会比较差。

为了克服K-means聚类方法的这个缺陷，考虑结合DBSCAN聚类算法和美国专利分类体系的优势来设计聚类步骤。首先，使用DBSCAN聚类算法按不同的年份对引文数据进行聚类，得到该数据集的聚簇类别数K1，然后结合美国专利商标局公布的大类数目为K2，经过计算可以得到聚簇的总类别数，最终取这两个类别数的平均值，即K=$(K1+K2)/2$，并向上取整。这样得到的这个K 就更加接近真实的类别数，然后将K值带入K-means聚类算法，按不同年份对引文数据进行聚类。经验证，上述处理可以较好的给出聚簇结果。

§3.2.3 类别标注

这个阶段相当重要，因为涉及到标注聚类结果的类标签，因为本文主要的研究内容是对新兴技术进行识别，那就意味着聚类结果的类标签只要分为新兴技术和非新兴技术两类就可以了，因此可以把它看成一个二分类问题。对于特定的技术，在现有的分类体系中为它们创建一个新的类别是衡量它们独立性的方法之一。如果某类技术的特征表明它们在未来几年内可以产生一个新的技术分类，那就将可以将该类技术标注为新兴技术。由此类推，对于一个给定的年份T，如果聚簇C_x是与T+1年的某类新兴技术相似的，可以类似的把C_x 标注为新兴技术，这意味着新兴技术相对于普通技术而言，可能会存在一些特征，为了更好的完成聚簇类别的标注任务，文中提出了一个聚簇类别标注算法，算法大致步骤如下，算法流程见图3.1。

算法步骤：

1、将T+1年公布的每一项专利依据主分类号来进行分组，将分组记为G_y；

2、如果该主分类号是在第T+1内新创建的，把G_y标注为新兴技术分组，否则标注为非新兴技术分组；

3、对第T年中的每一项专利，根据引文特征向量进行聚类,将聚簇记为C_x；

4、对于T年中任一聚簇C'_x计算与T+1年中所有分组G_y之间引文耦合相似度；

5、找到与聚簇C'_x引文耦合相似度最高的分组G'_y；

6、如果G'_y为新兴技术分组，将聚簇C'_x标为新兴技术，否则标为非新兴技术；

7、循环步骤4，直至T年中所有的聚簇C_x被标注完毕；

8、循环步骤1，直至所有需要处理年份中的专利都已完成聚类和标注。

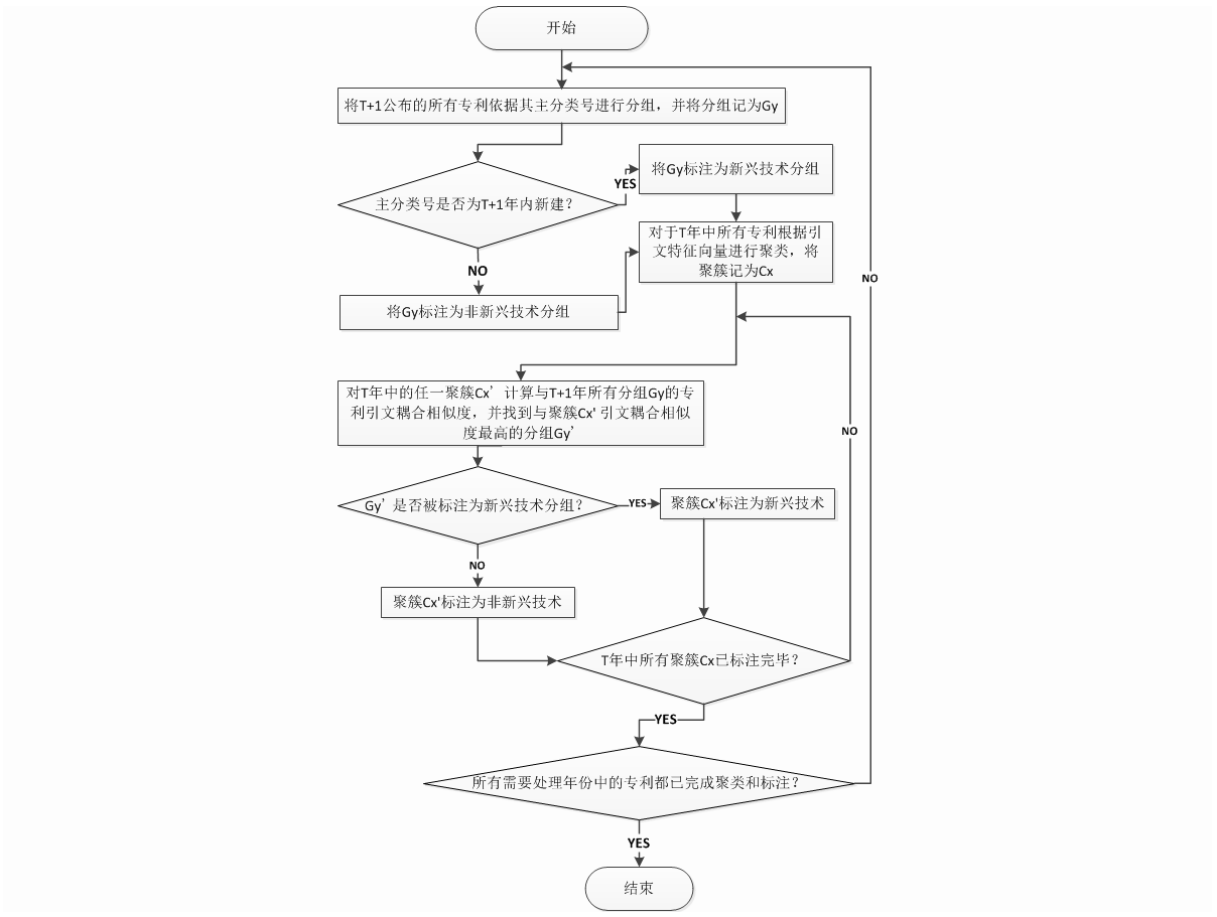

图 3.1 聚簇类标签标注算法流程图

在上述算法中，涉及到某一年新增的专利主分类号数据和文献耦合相似度的指标，下面给出相关说明：

对于美国在1975年至2009年之间每一年所有新增专利主分类号可以从美国专利商标局的官网[58]上获取，其中部分年份的具体统计数据见表3.2 所示：

表 3.2 不同年份新兴技术和非新兴技术类别数量

Year	ETCN	NETCN	NNCN	ETCMN	NETCMN
1987	6	490	5	131	149
1988	6	492	4	120	142
1989	1	491	5	68	177
1990	10	495	8	140	168
1991	0	484	10	-	180
1992	12	493	9	150	185
1993	20	479	8	157	188
1994	5	476	2	143	196
1995	9	486	7	170	194
1996	19	480	7	173	213
1997	0	468	3	-	150
1998	19	493	7	260	281
1999	0	463	4	-	295
2000	0	486	0	-	304
2001	2	483	1	282	314
2002	4	475	3	314	330
2003	0	468	2	-	331
2004	0	482	2	-	318
2005	0	475	0	-	283
2006	0	473	1	-	341
2007	0	475	0	-	308
2008	1	478	1	235	305

其中Year、ETCN、NETCN、NNCN、ETCMN、NETCMN分别指：年份、新兴技术聚簇数、非新兴技术聚簇数、下一年新建类别数、新兴技术簇平均数、非新兴技术簇平均数。

对于文献耦合相似度而言，一般情况下，如果两个对象引用了相同的对象，就认为这两个对象存在文献耦合的关系。这有点类似共引，但又与共引不同，因为文献耦合是可以被计算的。本文将文献耦合的观点拓展应用到专利聚簇中，因

此聚簇x和y的文献耦合相似度(BCS)就可以通过专利聚簇中专利的同引数量来计算。具体的计算公式为：

$$BCS_{XY} = \frac{n(C_x \bigcap C_y)}{n(C_x \bigcup C_y)} \quad (3\text{-}8)$$

这里BCS_{XY}是聚簇x和y的BCS值, C_x和C_y代表聚簇x和y中的专利集合。当专利引文数据集经过以上标注算法进行预处理之后，它们便可以转换成带有新兴技术和非新兴技术类别的标注数据集。

为了验证数据预处理的效果，针对聚簇类别标注算法给出一个评估策略，即根据类别标注结果的整体相似性来衡量算法的性能。通过计算每个簇前100个关键词的项频率- 逆文档频率（TF-IDF）来检验簇标注算法的性能。对于每个聚簇，分别创建由前100个特征项TF-IDF值所构成的向量Vi；此外，通过分析在给定年度期间出现的具有代表性的新技术类别，并从中提取前100 个特征项的TF-IDF值来创建另一个新兴技术向量Vj。然后计算由聚簇类别标注算法所标注出的具有最高引文耦合相似度的新兴技术聚簇所得到的向量Vi与同一年经过分析得出来的与之对应的新兴技术分组向量Vj之间的余弦相似性; 如果计算所得的余弦相似度值超过一个预先给定的阈值0.8，就可以说明类别标注算法所标注出来的聚簇与真实新兴技术类别组之间是相似的。在这项测试中，通过随机选定5个当时的新兴技术聚簇样本，并分别计算这5个聚簇与类别标注算法标注的新兴技术聚簇向量之间的余玄相似度，每个选定的聚簇和与之对应的新兴技术类别组向量之间的余弦相似性计算结果如表3.3 所示：

表 3.3 同一时间内标注的新兴技术聚簇和当时新兴技术之间的余玄相似度

新兴技术聚簇名称	新兴技术年份	与聚簇最相似的新兴技术	余玄相似度
cluster336-1989	1989	World Wide Web	0.93
cluster387-1989	1989	Digital TV	0.82
cluster462-1994	1994	DVD	0.89
cluster93-1995	1995	USB	0.72
cluster303-2004	2004	Flat Screens & HDTV	0.85

根据表3.3的结果，可以看到测试结果中，被标注算法标注为新兴技术的聚簇里面有80%的簇与其对应时间上的新技术之间的余弦相似性会超过0.8，这充分说明聚簇类别标注算法对新兴技术聚簇类别的标注工作取得了良好的效果。

最后，根据人为经验，对3.1.1中提出的所有特征根据OneR、相关性、Relief属性三个指标进行了重要性排名，以确定用以区分新兴技术聚簇和非新兴技术聚簇最显著的特征，一般情况下，专利类别是最重要的整体特征，其次是TCT，引用数量等。然而，有些特征与技术领域有关，在不同类别之间变化显著，例如物理科学中的TCT通常比非常快的技术领域更长。根据以上分析，可以得到按照不同指标进行衡量的特征排名情况，结果如表3.4所示：

表 3.4 特征重要性排名

特征项	OneR rank	相关性rank	Relief rank	平均排名
Patent Class	1	1	1	1.0
TCT	2	2	3	2.3
NumCitations	3	6	2	3.7
NonPatCitations	5	3	6	4.7
CASI	4	4	7	5.0
NumClaims	6	5	5	5.3
CTSI	7	7	4	6.0

§3.3 基于深度学习的新兴技术识别算法

基于深度学习理论的新兴技术识别算法的核心是深度神经网络层的数据重构算法，为了更好的将DBNs算法引入基于深度学习的新兴技术识别模型中，从系统角度进行分析、设计并实现了整个深度学习的过程。图3.2为文中描述的深度学习系统的体系结构图。

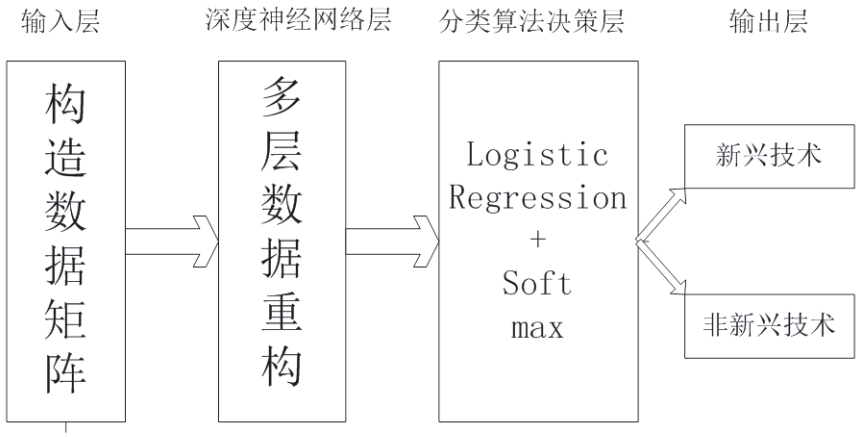

图 3.2 深度学习系统体系结构图

如图3.2所示，本文所构造的基于深度学习的新兴技术识别算法一共分为四个层次，首先是输入层，本层需要对输入数据进行预处理，形成统一格式的数据矩阵；然后就是深度神经网络层，该层由3层RBM堆叠而成，主要功能就是对数据进行重构，自动提取出合适的特征；接下来是分类器所在的决策层，该层使用Logistic Regression算法来设计分类器，然后再对分类结果应用Softmax算法进行概率转换。将结果中概率较大的所对应的下标作为分类结果，因为原分类结果只有两个维度，因此最终的分类结果只有0或者1，0代表非新兴技术，1代表新兴技术。

§3.3.1 构造数据矩阵

经过3.2的预处理步骤之后，得到数据集中一共包含8个特征属性和一个对应的类标签。为便于本文所设计的系统可以对所有样本进行处理，本文约定数据格式为只包含Num Claims,Num Citations,Non-Pat Citations,TCT,CTSI, CASI 6项属性的统一格式，本文通过将数据预处理后得到的数据构造成数据矩阵，将这个矩阵的结构形式作为新兴技术识别算法的输入层读入数据的统一格式，如图3.3所示。

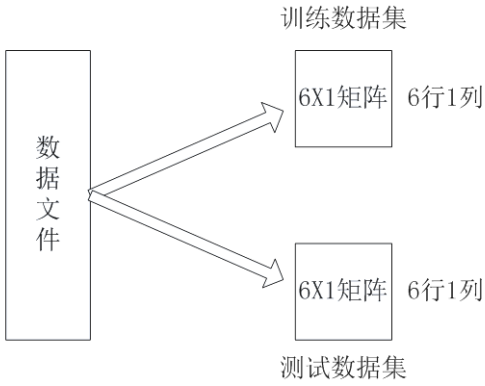

图 3.3 用于深度学习模型的数据矩阵

图3.3中展示的矩阵是一个针对600个聚簇的训练样本所构建的数据矩阵。对于这600个样本，，首先将它们分成了用于训练和测试的2个数据集，训练数据有480个实例，测试数据包含120 个实例。每一个样本以上面所选的六个特征中的每一个特征分别作为一行，来构造6行1列的数据矩阵，并且把标签项的类别做相应的处理，使得new映射成1,old映射成0以便用于后续步骤的处理，即训练集是一个480 x 6 x 1的高维矩阵，测试集是一个120 x 6 x 1的矩阵。其中应该注意的是在图3.3中所述的数据文件指的是经过预处理后所得到的数据文件。

§3.3.2 重构算法的选取以及参数学习

①重构算法的选取

在深度神经网络层，本文选用了深度信念网络(DBNs)来构建深层网络。在这个模块，需要使用相应的重构算法并设计合适的网络结构才能实现对输入数据的合理重构。在深度学习中，一般常用的重构算法是RBM和DA，针对本文所使用的数据集，经过实验得出如下结果，见图3.4重构算法在不同迭代次数下的重构误差变化图和图3.5隐藏层神经元个数与重构误差的关联关系图。

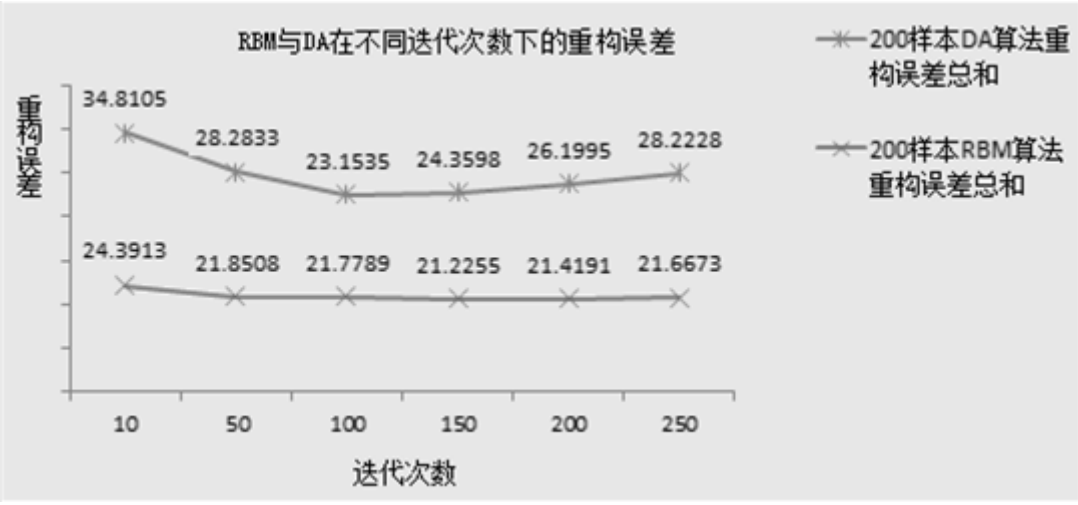

图 3.4 重构算法在不同迭代次数下的重构误差对比图

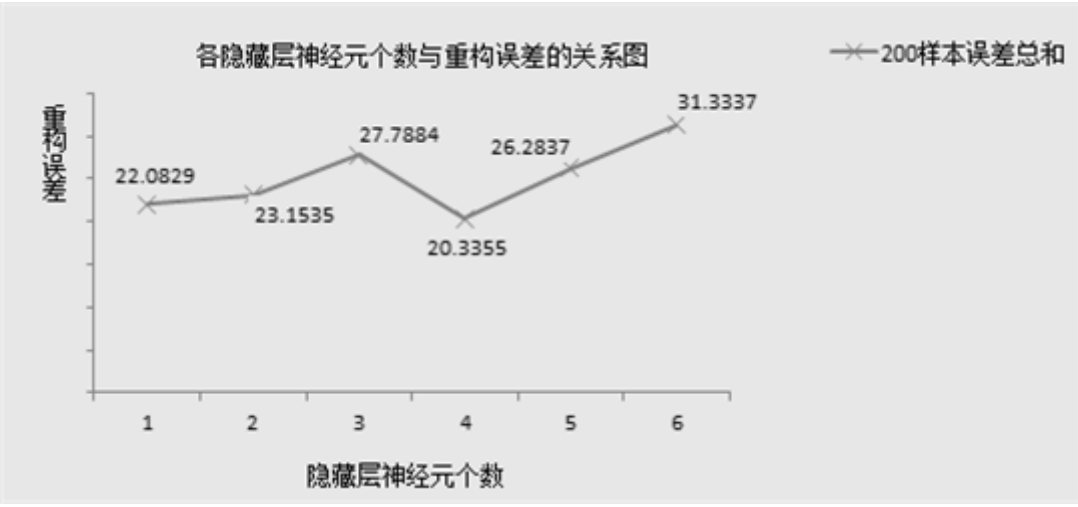

图 3.5 隐藏层节点数量与重构误差之间的关联关系图

如图3.4所示，通过在200个实验样本下的实验结果可以直观的看出，在重构算法的表现上，RBM比DA的重构误差要更小而且更稳定。因此本文选取RBM算法作为深度信念网络各层之间的重构算法。如图3.5所示，在200个实验样本下，重构误差与隐藏层神经元个数设置有一定的关系，实验结果表明当把隐藏层神经元个数设置为4个时，重构误差达到最小，因此，本文将深度神经网络层中的隐藏层神经元个数设置为4个。

②RBM调节

RBM的调节过程是深度神经网络层的核心，也是整个深度学习最为重要的一个阶段，它是以逐层递进的方式，利用受限制的玻尔兹曼机算法来调节数据重构阶段相邻两层网络之间的节点的权值。根据分析结果可知，RBM的调节阶段

是深度学习模型的核心部分，在传统神经网络的参数训练中，如何选择合适的初始化参数来赋值给整个神经网络是一个难点，一旦参数选择不佳，往往会导致模型效果较差。

图3.6展示了整个深度信念网络里面各层之间RBM调节的详细过程。最初由第一可见层向第一隐藏层进行转化，经历本次转化以后，以第一隐藏层为标准进行抽样，得到隐藏层各个节点的状态，再反向由第二隐藏层向第二可见层转化，这次转化结束以后，完成最后一轮由第三可见层到第三隐藏层的转化。要进行三次转化的原因是为RBM内部的参数调节提供训练目标，通过降低重构矩阵与原矩阵的差异来达到调节RBM参数的最终目标。

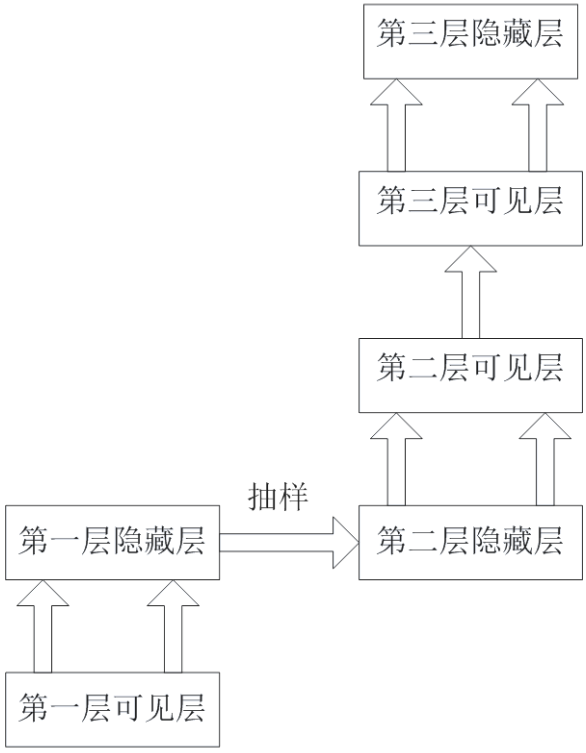

图 3.6 深度神经网络各层之间RBM调节详细过程图

基于图3.6所描述的RBM调节的详细过程，接下来从算法实现的角度对RBM的调节过程进行具体描述。

RBM调节算法：RBM（W,b,c,v_0）

W是RBM层与层之间的链接权重矩阵

b是RBM隐藏层的偏置量

c是RBM输入层的偏置

v_0是RBM训练样本集合中的一个样本点。

对于所有隐藏层的神经元i：

计算$Q(H_{0i}=1|v_0)$,即进行层与层之间的映射运算$sigm(b_i+\sum_j W_{ij}*v_{0j})$

从$Q(H_{0i}=1|v_0)$中进行抽样得到H_{0i}

对于所有可见层神经元j：

计算$P(v_{1j}=1|H_0)$，即进行层与层之间的映射运算$sigm(c_j+\sum_i W_{ij}*H_{0i})$

从$P(v_{1j}=1|H_0)$中进行抽样得到v_{1j}

对于所用隐藏层的神经元节点i：

计算$Q(H_{1i}=1|v_1)$，即进行层与层之间的映射运算$sigm(b_i+\sum_j W_{ij}*v_{1j})$

最后更新链接的偏置参数的权值：

$$W=W+\varepsilon(H_0v_0^{'}-Q(H_1=1|v_1)v_1^{'})$$

$$b=b+\varepsilon(H_0-Q(H_1=1|v_1))$$

$$c=c+\varepsilon(v_0-v_1))$$

③RBM参数学习算法

根据第二章的2.3节可知，学习RBM的核心任务就是求出模型中参数θ的值，以便用于拟合给定的学习样本，本文使用对数似然度极大化的思想来获取RBM算法中参数θ，θ的表达式定义如下：

$$\theta^*=argmax_\theta\sum_{k=1}^{K}\log P(v^{(k)}|\theta) \tag{3-9}$$

为了获得最优参数，可以使用随机梯度上升法进行求解，其中最关键的步骤是计算关于各个模型参数的偏导数。由于

$$L(\theta)=\sum_{k=1}^{K}\log P(v^{(k)}|\theta)=\sum_{k=1}^{K}\log\sum_h P(v^{(k)},h|\theta) \tag{3 -10}$$

$$=\sum_{k=1}^{K}\log\frac{\sum_h e^{[-E(v^{(k)},h|\theta)]}}{\sum_v\sum_h e^{[-E(v,h|\theta)]}}$$

$$=\sum_{k=1}^{K}[(\log\sum_h e^{[-E(v^{(k)},h|\theta)]})-(\log\sum_v\sum_h e^{[-E(v,h|\theta)]})] \tag{3-11}$$

则对数似然函数关于 θ 的梯度为：

$$\frac{\partial L}{\partial\theta}=\sum_{k=1}^{K}\frac{\partial}{\partial\theta}((\log\sum_h e^{[-E(v^{(k)},h|\theta)]})-(\log\sum_v\sum_h e^{[-E(v,h|\theta)]})) \tag{3-12}$$

$$=\sum_{k=1}^{K}(\sum_h\frac{e^{[-E(v^{(k)},h|\theta)]}}{\sum_h e^{[-E(v^{(k)},h|\theta)]}}\times\frac{\partial e^{[-E(v^{(k)},h|\theta)]}}{\partial\theta}-\sum_v\sum_h\frac{e^{[-E(v,h|\theta)]}}{\sum_v\sum_h e^{[-E(v,h|\theta)]}}\times\frac{\partial e^{[-E(v,h|\theta)]}}{\partial\theta})$$

$$=\sum_{k=1}^{K}(\langle\frac{\partial(-E(v^{(k)},h|\theta))}{\partial\theta}\rangle_{P(h|v^{(k)},\theta)}-\langle\frac{\partial(-E(v,h|\theta))}{\partial\theta}\rangle_{P(h|v,\theta)}) \tag{3-13}$$

上式$\langle\cdot\rangle$P表示求关于分布P的均值。假设使用D和M来简记$P(h|v^{(k)},\theta)$和$P(v,h|\theta)$这两个概率分布，则对数似然函数关于连接矩阵W_{ij},a_i,b_j,的偏导数分别为:

$$\frac{\partial \log P(v|\theta)}{\partial w_{ij}} = \langle v_i * h_j \rangle_D - \langle v_i * h_j \rangle_M, \tag{3-14}$$

$$\frac{\partial \log P(v|\theta)}{\partial a_i} = \langle v_i \rangle_D - \langle v_i \rangle_M, \tag{3-15}$$

$$\frac{\partial \log P(v|\theta)}{\partial b_j} = \langle h_j \rangle_D - \langle h_j \rangle_M, \tag{3-16}$$

§3.3.3 深度学习模型的反馈微调

深度模型的反馈微调过程主要通过三个步骤来完成：加载参数、构造数据矩阵、循环微调。其中前面两个过程是整个深度模型的前期准备工作，而循环调节过程才是核心。图3.7描述了文中设计模型的基本架构与流程：

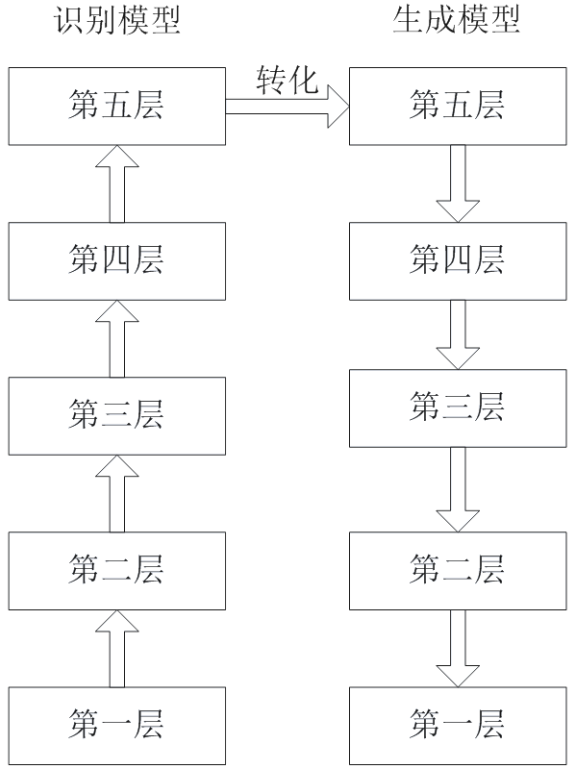

图 3.7 深度学习模型的整体反馈微调流程图

(1)加载参数。在图3.2所描述的深度学习系统体系结构中，在执行深度模型整体微调之前，需要完成深度神经网络的重构任务，在重构过程中，一共经历了4次RBM调节，经过这4次调节，整个深度模型的层与层之间便拥有了自己的初始连接权值。加载参数这个过程的主要任务就是将重构过程已经训练好的参数的权值加载到整个模型中作为模型参数的初始值。

(2)构造数据矩阵。这一过程与3.4.1节所述的构造数据矩阵的过程完全一致，在此阶段重新构造数据矩阵的原因是经历了多次RBM的调节以后，原来所构建的数据矩阵中的原始数据会遭到破坏。

(3)循环微调。这个阶段是整个模型的核心部分，本文结合了自底向上的识别模型和自顶向下的生成模型对系统进行微调。经过各层次的逐步优化，生成模型就可以重构出具有较低误差的训练样本，通过以上步骤模型可以自动学习出原样本的数据特征，即最高层次的抽象表示形式。

§3.4 本章小结

本章主要内容是围绕新兴技术识别算法的构建来展开的，共分为三个小部分。第一部分，经过深层分析，完成了特征选取的任务，一共选取了七项主要特征并给出了相应的定义；第二部分，本文通过对专利引文数据进行索引创建、并对其进行聚类与标注，使数据达到实验可用；第三部分，是基于深度学习的新兴技术识别算法，本文利用深度学习的多层RBM算法来构建识别模型，然后对数据矩阵的构造和RBM的训练过程和参数学习算法以及整个深度模型的反馈微调进行了描述。

第四章　算法评估准则及实验对比分析

§4.1　算法评估准则

本文主要采用算法的重构误差和识别准确率F1值等指标来对基于深度学习的识别算法的性能进行评估。其中重构误差指标，可以用来衡量深度学习阶段特征重构的效果，即衡量输入对象最高层次的抽象表示形式，这是一个用来衡量深度学习阶段重构质量的指标。而识别准确率和F1值，它可以直接反映整个基于深度学习的新兴技术识别算法的整体性能的指标。在识别问题中，模型的识别准确率和F1值需要根据算法的实验结果所给出的混淆矩阵来得出。

①重构误差

在深度学习阶段，需要进行多次的RBM重构，这样得到的重构对象的结果与输入对象的结果会存在一定的误差。重构误差越小，就说明重构效果越好，所提取特征的最高层次的抽象表示形式就越接近真实的本质特征；若重构误差太大，就说明重构效果比较差，对提取特征最高层次的抽象表示形式不理想。

重构误差的计算并不复杂，对RBM 的重构进行简单的评估是有实际意义的。下面给出如何计算重构误差的步骤：

初始化误差：Error=0

循环(所有$v^k, k \in 1, 2, \cdots, T$):

依次求解条件概率分布$P(h, v|\theta)$,从条件概率分布中抽取$h \in \{0, 1\}$；

依次求解条件概率分布$P(v', h|\theta)$,从条件概率分布中抽取$v' \in \{0, 1\}$；

$Error = Error + |v' - v^k|$；

返回总的重构误差Error。

②混淆矩阵及相关指标的定义

在二分类评估中，实例的分类结果只会被判定为下列四种结果之一：TP、FP、FN、TN。根据这四种类型，下面给出混淆矩阵的一般结构形式：如表4.1所示:

表 4.1 混淆矩阵的结构

混淆矩阵	预测正实例	预测负实例
真实正实例	TP	FN
真实负实例	FP	TN

根据混淆矩阵的结构对准确率、查准率、查全率和F1值分别给出如下定义：

准确率 =正确预测的正反例总数/测试总数

$$= \frac{(TP+TN)}{(TP+FN+FP+TN)} \tag{4-1}$$

上述正实例对应新兴技术，新兴技术评估指标的计算公式：

$$P_{new} = \frac{TP}{(TP+FP)}, R_{new} = \frac{TP}{(TP+FN)}, F_{1,new} = \frac{2*P_{new}*R_{new}}{(P_{new}+R_{new})} \tag{4-2}$$

上述负实例对应非新兴技术，非新兴技术评估指标的计算公式：

$$P_{old} = \frac{TN}{(TN+FN)}, R_{old} = \frac{TN}{(TN+FP)}, F_{1,old} = \frac{2*P_{old}*R_{old}}{(P_{old}+R_{old})} \tag{4-3}$$

§4.2 数据集的获取

众所周知，大多数算法的训练及验证离不开对数据的依赖，本文所述的基于深度学习的新兴技术识别算法也一样，从某种程度上来说，数据的完整性、可靠性和稳定性就可以用来作为一个衡量模型性能的基础指标。除了专利数据库之外，会议录和社会网络分析等也是提取新兴技术识别信息的主要数据资源[57]。然而利用专利数据库可以对授权的新技术得出最优的投资组合。结合本文研究内容的特性，经过综合分析，决定通过以下两种方式来获取相关实验所需的数据集。

§4.2.1 网上收集数据

通过对中国专利检索数据库、日本专利特许厅、欧洲专利局和美国专利商标局等几个全球主要的专利数据提供商的引文数据进行对比分析后，得出美国专利商标局所提供引文数据相对来说比较完整，因此在美国专利商标局的官网上收集了自1975-2009年已授权专利的引文信息，其中包括发明专利、外观设计专利的引文数据。

§4.2.2 利用网络爬虫获取数据

通过对上述所收集的引文数据进行解读与分析，发现该数据只是涵盖了1975-2009年之间授权的全部专利引文信息，而并未包含1975年之前的授权专利引文

信息。例如，专利A授权于2008，它总共有25条引文信息记录，其中有20条完整的引文信息，即引用对象的授权时间是处于1975-2009年之间，而还有5条引文记录的引用对象授权时间是在1975年之前，那么这5条引文记录的信息就缺失了，这就导致数据的完整性受到了破坏，为了保证数据的完整性，借助Google专利搜索数据库来完善缺失的引文信息。利用Java语言编写的网络爬虫开源工具包Jsoup从Google专利搜索系统中爬取上面获得的数据集中被引用专利授权时间在1975年之前的引文记录信息，然后把得到的数据更新到之前的数据集中，以形成完整的引文数据集。

通过以上两种方式就可以获取到实验所需专利引文的基础数据，上述基础数据经过3.2节所述的数据预处理步骤之后，可得到相关实验所需的数据集。数据集结构见表4.2所示（部分数据）：

表 4.2 预处理后的数据集结构形式（部分数据）

CN	PC	NC	Nc	NPC	TCT	CTSI	CASI	真实类别
c24-2004	347	0.8781	0.3033	0.0086	0.2694	0.0201	0.1636	new
c435-2004	264	0.5953	0.6854	0.0291	0.3708	0.0078	0.0924	new
c112-2004	101	0.8073	0.4763	0.0096	0.2754	0.0116	0.0815	new
c417-2004	600	0.6863	0.6574	0.0841	0.2198	0.0110	0.0529	new
c388-2004	705	0.3945	0.8345	0.1072	0.2445	0.0083	0.0457	new
c387-1988	250	0.6851	0.5654	0.0302	0.3679	0.0278	0.4873	old
c243-1988	200	0.8844	0.3388	0.0048	0.2568	0.0190	0.1102	old
c419-1988	110	0.5706	0.6716	0.0041	0.3870	0.0202	0.0631	old
c391-1988	426	0.6802	0.6149	0.0491	0.3364	0.0269	0.0255	old
c366-1988	208	0.5854	0.6823	0.0251	0.2770	0.0154	0.0891	old

表4.2中属性名CN,PC,NC,Nc,NPC,TCT,CTSI,CASI分别表示：聚簇名称，专利分类号，权利要求项数，专利引文数量，非专利引文数量，技术生命周期，被引技术的相似性指数，被引技术所有者平均相似性指数。

§4.3 实验结果及对比分析

如果将新兴技术识别问题看成一个二分类的问题，那么就可以使用其它的机器学习方法来做分类预测，然后通过分析其他方法得到的实验结果就可以与现有算法形成对比，这样可以更好的反映现有算法的优缺点。因为现有模型采用了识别准确率来评估模型的性能，因此为了形成更好的对比结果，要求在其它的分类方法中也必须要采用这个指标来评估模型的性能。机器学习的分类方法有很多，这里只是选取了部分经典分类方法的实验结果来形成对比分析，最后再给出这个基于深度学习的新兴技术识别算法和传统机器学习算法在实验结果上的简单对比。

§4.3.1 深度学习算法的训练

深度学习算法的参数训练主要包括两个任务，分别为RBM调节和深度学习识别算法整体的反馈微调过程。

1.RBM的训练与分析

RBM的训练是在整个深度学习算法的构建初期所需要完成的一项任务。对它的训练是一个反复转换的过程。实验的主要目的是通过不同实验结果进行对比分析RBM算法在不同循环次数下的重构质量，从而选出最优的迭代次数。图4.1和图4.2分别是RBM算法循环了1次和5次所得到的重构结果，从图中可以清晰的得出结论：迭代5次得到的结果比迭代1次的结果要准确。通过实验结果的对比可知，图4.2中的重构数据与测试数据走向更加一致，而图4.1中两者之间的波动幅度非常大，这说明重构效果会受迭代次数的影响。

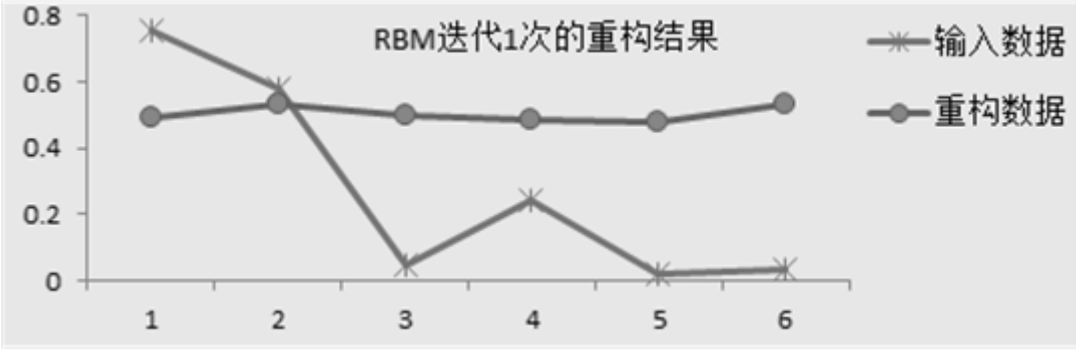

图 4.1 RBM迭代1次的实验结果

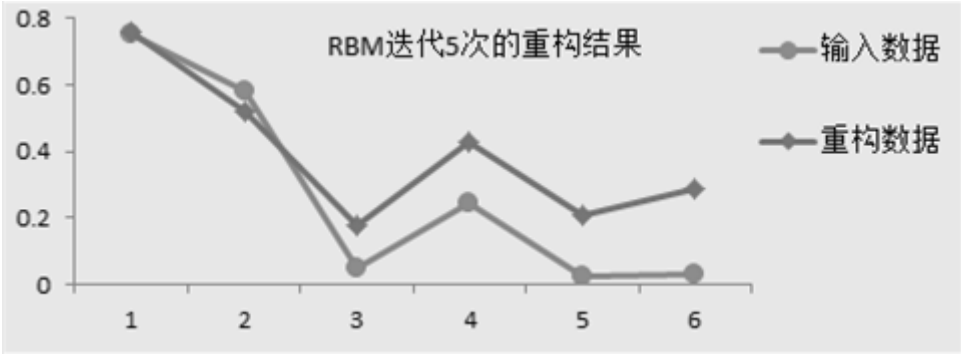

图 4.2 RBM迭代5次的实验结果

图4.3，分别表示RBM算法经过10次迭代、30次迭代、50次迭代和100次迭代所得到的重构结果，从图4.3中可以发现在迭代10次的时候，重构数据与测试数据的走势基本吻合，迭代次数达到30次的时候，重构数据在某些维度上出现了较大幅度的波动，迭代50次和迭代100次所得到的重构结果几乎没有变化。

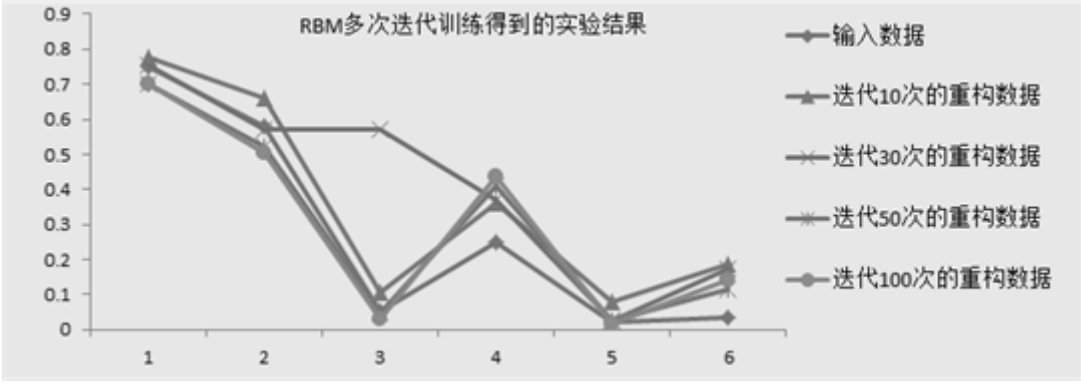

图 4.3 RBM迭代多次的实验结果

通过以上试验结果的对比，可以发现本文所选择的RBM算法的重构能力和迭代次数之间并不是呈现着简单的递增或者递减的关系。在最初的时候，随着迭代次数的上升，重构效果会变好，但是随着迭代次数达到某个阀值时，迭代次数与RBM重构效果呈现负相关。最终在本文的深度学习算法中采用50次迭代来对RBM进行调节。

2.深度学习反馈微调的实验与分析

在深度学习算法的实验阶段，本文首先对深度学习算法的重构能力进行了实验测试。深度学习算法通过识别模型得到自底而上的抽象表示形式，然后利用生成模型生成目标对象抽象层的表示，经过多次微调之后，使重构结果逼近于原目标，从而将重构误差降低到可接受的范围之内。

图4.4是输入数据经历自底向上识别模型的深度抽象过程，然后又经过了自顶而下的生成模型的深度学习过程。从图中可以看出，第1次深度学习的数据还不是太理想，但是基本的趋势还是一致。

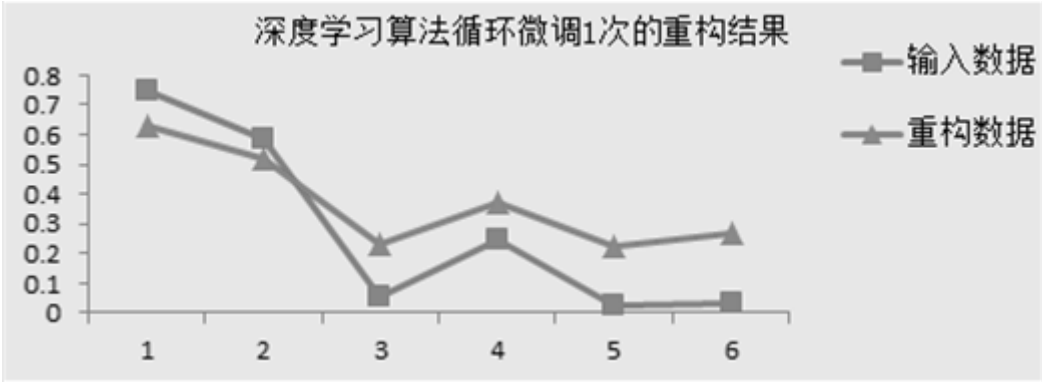

图 4.4 深度学习模型迭代1次的实验结果

图4.5分别是算法在迭代100、200、500和1000次的实验结果对比情况。经过以上实验结果的对比发现，开始时伴随微调次数的不断增加，被重构出来的结果就越接近输入数据，当迭代次数达到100次的时候，重构出来的结果和原始输入数据的误差就已经非常小了，这充分说明了深度信念网络在本文中表现出色。

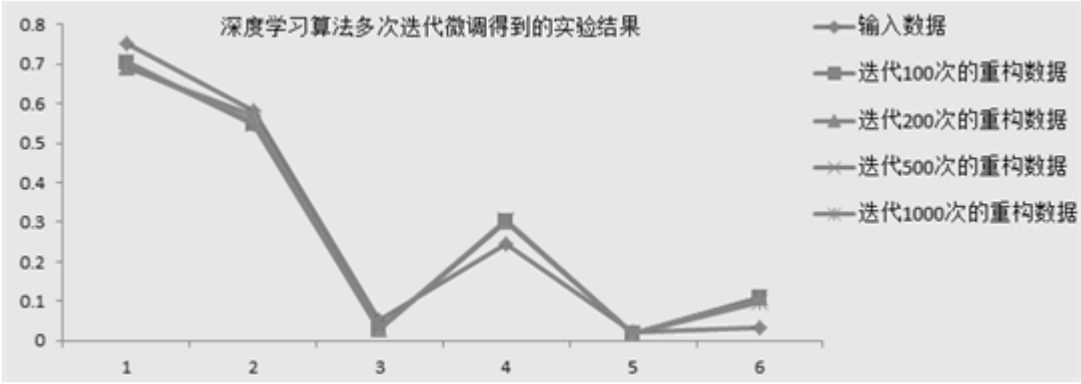

图 4.5 深度学习模型迭代多次的实验结果

图4.6和图4.7描述的是在深度学习过程中，整个深度学习系统的重构误差和交叉熵的变化情况。从图中可以看出，随着迭代次数的增大，重构误差在逐渐下降，在迭代前期，重构误差下降非常快，越到后面，下降幅度越小，但是依然保持着微小的变化。交叉熵在迭代初始阶段下降极快，但当迭代次数达到30次的时候，交叉熵就趋于稳定，这表明此时的深度学习系统已经接近最佳学习状态即已经找到各节点参数的最优权值。

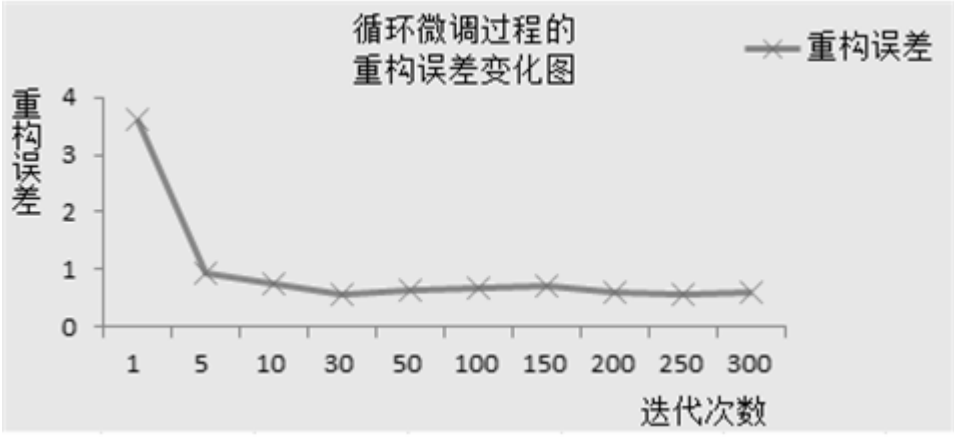

图 4.6 深度学习系统的重构误差变化图

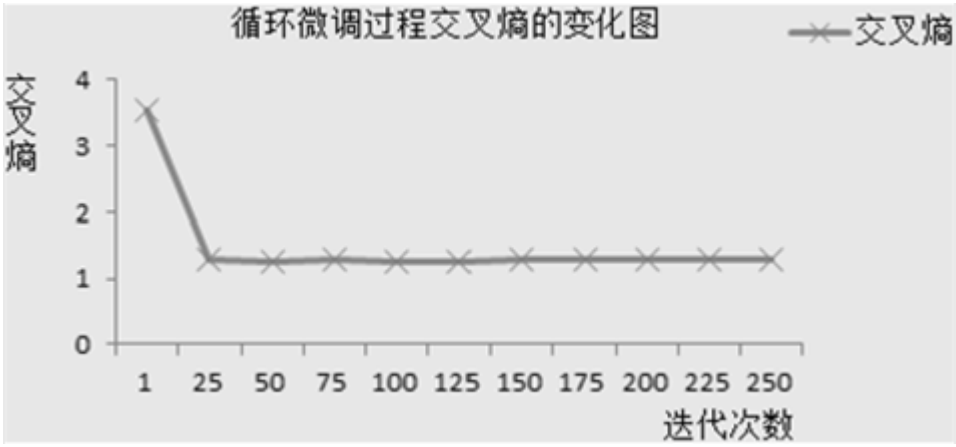

图 4.7 深度学习系统的交叉熵变化图

§4.3.2 基于深度学习的新兴技术识别算法的实验结果

为了检验基于深度学习的新兴技术识别算法的性能，需要利用经过预处理的数据集来验证模型的性能。根据第三章的介绍，本实验所使用的数据集是经过对美国专利商标局1975年-2009 年之间已授权专利的引文数据进行预处理之后得到的带类别标签的数据集，该测试数据集的部分数据如表4.3所示。

表 4.3 实验所使用的测试数据集（部分数据）

CN	PC	NC	Nc	NPC	TCT	CTSI	CASI	真实类别
C114-1988	73	0.8444	0.4144	0.0225	0.2982	0.0225	0.0193	1
C254-1988	200	0.5832	0.7280	0.0060	0.3180	0.0273	0.1595	1
C387-1988	250	0.6851	0.5654	0.0302	0.3679	0.0278	0.4873	0
C63-1988	525	0.8840	0.2497	0.0133	0.3184	0.0173	0.0418	1
C233-1988	74	0.8697	0.3635	0.0062	0.2729	0.0155	0.0857	1
C419-1988	110	0.5706	0.6716	0.0041	0.3870	0.0202	0.0631	0
C366-1988	208	0.5854	0.6823	0.0251	0.2770	0.0154	0.0891	0
C466-1988	526	0.3902	0.6899	0.0359	0.5639	0.0315	0.6058	0
C166-1988	123	0.6952	0.5880	0.0043	0.2786	0.0384	0.6452	1
C98-1988	324	0.5223	0.7164	0.3169	0.2447	0.0274	0.0580	0

算法在以上测试数据集上进行实验所得的结果如表4.4所示：

表 4.4 在测试数据集上实验所得结果（部分数据）

CN	PC	NC	Nc	NPC	TCT	CTSI	CASI	预测类别
C114-1988	73	0.8444	0.4144	0.0225	0.2982	0.0225	0.0193	1
C254-1988	200	0.5832	0.7280	0.0060	0.3180	0.0273	0.1595	0
C387-1988	250	0.6851	0.5654	0.0302	0.3679	0.0278	0.4873	0
C63-1988	525	0.8840	0.2497	0.0133	0.3184	0.0173	0.0418	1
C233-1988	74	0.8697	0.3635	0.0062	0.2729	0.0155	0.0857	1
C419-1988	110	0.5706	0.6716	0.0041	0.3870	0.0202	0.0631	0
C366-1988	208	0.5854	0.6823	0.0251	0.2770	0.0154	0.0891	1
C466-1988	526	0.3902	0.6899	0.0359	0.5639	0.0315	0.6058	0
C166-1988	123	0.6952	0.5880	0.0043	0.2786	0.0384	0.6452	1
C98-1988	324	0.5223	0.7164	0.3169	0.2447	0.0274	0.0580	1

表4.3和表4.4中属性名CN,PC,NC,Nc,NPC,TCT,CTSI,CASI分别表示：聚簇名称，专利分类号，权利要求项数，专利引文数量，非专利引文数量，技术生命周期，被引技术的相似性指数，被引技术所有者平均相似性指数。

本实验里面数据集的划分准则采用80%作为训练集，20%作为测试集，算法设定学习率的初始值为0.01。部分测试集的实验结果如上表4.3所示。

表4.3和表4.4中最后一列的值分别表示聚簇实际标注类别和算法预测的类别（1代表新兴技术类别，0代表非新兴技术类别）。根据以上实验结果可知，测试集合总共有120个样本，预测正确的实例为86个，预测错误的实例为34个，因此通过计算可以得到这个基于深度学习的新兴技术识别算法的识别准确率为71.667%。

§4.3.3 新兴技术识别算法与其它经典算法的简单对比

通过将新兴技术的识别问题看成一个二分类的问题，就可以使用其它经典的浅层机器学习算法来做简单的分类与预测。这里采用600个样本数据中的80%作为训练集，20%作为测试集进行一次随机实验得出相应结果，本次试验分别使用单分类器和组合分类器进行试验，从而形成相互对比；其中单分类器包括逻辑斯

蒂克回归,径向基神经网络,朴素贝叶斯算法,决策树,支持向量机等，组合分类器包括LogitBoost,AdaBoost,RandomForest,NBTree，以及文中提出的DBNS+Logistic算法，以上算法实验结果分别如表4.5和表4.6所示：

表 4.5 数据集按80%训练、20%测试时各单分类器所得实验结果

评估指标&算法名称	LR	J48	NB	RBF network	SVM
准确率（%）	59.75	49.992	66.459	66.458	70.618
R_{new}（%）	61.30	10.0	61.60	65.30	73.60
R_{old}（%）	58.30	89.70	71.30	67.60	67.60
P_{new}（%）	59.40	49.20	68.20	66.80	69.40
P_{new}（%）	60.10	50.0	65.10	66.20	72.0
$F_{1,new}$（%）	60.2	16.6	64.70	66.0	71.40
$F_{1,old}$（%）	59.2	64.2	68.10	66.90	69.80

表 4.6 数据集按80%训练、20%测试时各组合分类器所得实验结果

指标&算法名称	L-Boost	A-Boost	RF	NBT	DBNS+Logit
准确率（%）	61.154	66.459	49.922	69.267	71.667
R_{new}（%）	59.60	61.60	10.0	60.30	80.0
R_{old}（%）	65.40	71.30	89.70	78.20	63.33
P_{new}（%）	62.10	68.20	49.20	73.40	68.57
P_{new}（%）	60.30	65.10	50.0	66.40	76.0
$F_{1,new}$（%）	59.40	64.70	16.6	66.20	73.85
$F_{1,old}$（%）	62.80	68.10	64.2	71.80	69.09

表中的缩写：L-Boost、A-Boost、RF、NBT、DBNS+Logit分别是指：Logit-Boost、AdaBoost 、Random-Forest、NBTree、DBNS+Logistic算法

各算法实验结果如图4.8所示：

图中分别使用：准确率、新兴技术查全率、非新兴技术查全率、新兴技术查准率、非新兴技术查准率、新兴技术F1值、非新兴技术F1值等指标来评估算法的性能。

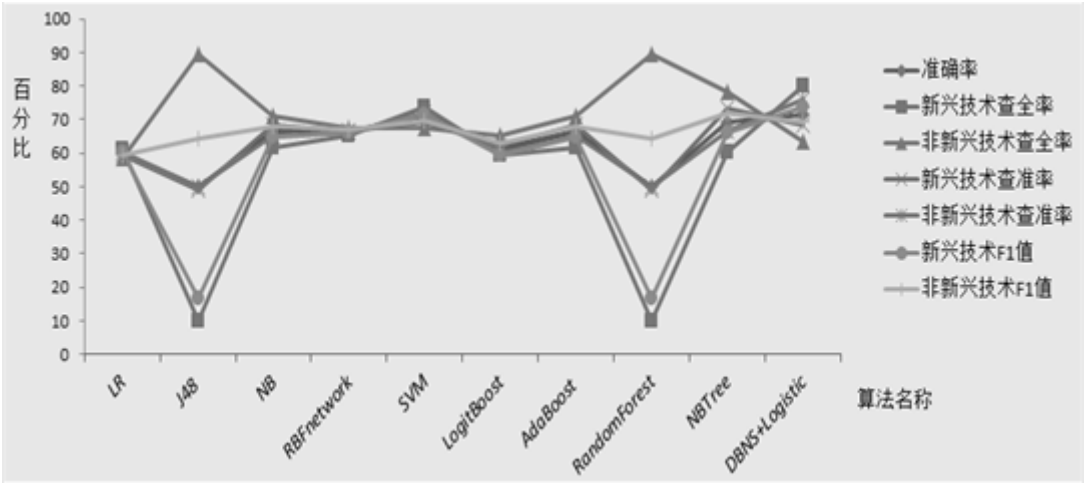

图 4.8 新兴技术识别算法与传统算法的实验结果对比图

根据表4.4和表4.5以及图4.8中的实验结果，可以看出在单分类器的测试中,支持向量机的准确率最高，可以达到70.618%，未结合深度置信网络的单个Logistic Regression算法的准确率只有59.75%，在传统组合分类器中，NBTree的准确率最高可达到69.267%，而使用基于深度信念网络加上Logistic回归的组合新兴技术识别算法的准确率比其他算法都要更高一点，可以达到71.667%。根据以上实验结果，可以得出如下结论：

1、文中所提出的基于深度学习的新兴技术识别算法在新兴技术识别上的准确率相对传统算法有一定的提升；

2、新算法的各项指标值相对来说比较稳定，而且F1值与其他算法相比显得更优，说明新算法在新兴技术识别问题上具有较好的适应性。

3、从实验结果来看，表3.4中所得到的特征重要性排名还是可以做进一步优化；因为根据人为经验选择的最优特征组合与采用深度学习自动提取的最优特征组合所得到的实验结果存在一定的差距，这进一步说明了新算法的优势。

§4.4 本章小结

本章的主要内容是以基于深度学习的新兴技术识别算法的评估及实验结果的对比分析为中心来展开。首先介绍了模型的评估准则，它包括了重构误差和准确率等评估指标；接下来，介绍了数据集的获取方式，本文采取了从网上下载已有的数据和利用网络爬虫这两种方式来获取数据集；然后对深度学习算法中的RBM重构实验结果和深度信念网络的整体循环微调的实验结果进行了详细分析；接下来给出了本文所构造的基于深度学习的新兴技术识别算法的实验结果；最后将文中所提出算法的实验结果与其它机器学习算法的实验结果进行了简单对比，并给出了相应的结论。

第五章　总结与展望

§5.1　全文总结

本文的研究主题是“基于引文分析和深度学习的新兴技术识别算法研究”，它属于深度学习和新兴技术识别领域的研究内容。因此，本文开篇就阐述了新兴技术的概念、特征及类型定义，并对近年来的研究现状进行了分析叙述。此外，还详细描述了相关的背景知识，例如引文分析。本文主要内容可分为研究综述、算法理论的介绍、数据预处理与基于深度学习识别算法的构建以及算法评估准则及实验对比分析四个部分。

首先，对新兴技术识别领域的现有方法进行了全面的综述，并揭示了它们的弊端，为后面的研究指明了方向。

接下来，对文中需要用到的算法原理以及它们的优缺点进行了简单介绍，对深度学习的相关知识进行了概述，并对其中的数学原理给出了推导过程；此外还介绍了用来构建新兴技术识别算法的受限玻尔兹曼机和逻辑斯蒂克回归算法。

然后，在特征处理上面给出了详细介绍，例如特征的选取、定义和相关计算公式。之后对深度学习中的深度信念网络的原理进行了简单的介绍，并根据深度信念网络结合逻辑斯蒂克回归算法构建出了基于深度学习的新兴技术识别算法。

最后，为了检验基于深度学习的新兴技术识别算法的性能，本文定义了一些评估指标，然后给出了数据集的获取方式，接下来给出了深度学习算法的训练和循环微调的实验结果以及文中所构建算法的实验结果，最后利用对比分析的思想，通过结合一些经典的机器学习算法的实验结果来形成简单的对比，最后得出了基于深度学习的新兴技术识别算法的识别准确率为71.667% 的结论。

§5.2 未来展望

深度学习算法的理论研究已经比较成熟，但是真正的应用还是相对较少。文中基于多层RBM的深度信念网络构建了一个新兴技术识别模型，未来可以进一步改进该模型，计划用深度卷积神经网络来构建深度模型并结合不同的分类算法来构建识别算法，这样可以扩展算法的适应性，以便用于更广泛的研究领域。

此外，文中所选用的数据集是来自美国专利商标局的专利引文数据，未来会考虑改进该模型的结构设计并将其迁移到其它相关的领域。

参考文献

[1] 周潇. **新兴技术热点领域识别及技术路线图研究** [D]. 北京：北京理工大学, 2015.

[2] George Day. **Wharton on Managing Emerging Technologies** [J].John Wiley & Sons, 2000.

[3] 赵振元,银路,成红. **新兴技术对传统管理的挑战和特殊市场开拓的思路** [J].中国软科学, 2004, **7**: 72-77.

[4] 华宏鸣. **从技术概念的要素分析谈高新技术** [J].研究与发展管理, 1995,(1):14-18.

[5] 银路. **新兴技术管理导论** [M]. 北京：科学出版社，2010.

[6] 佚名. **2012新兴技术展望** [J]. 商界:评论, 2012(4):29-29.

[7] 侯剑华,王鹏. **国内新兴技术及其管理研究综述** [J]. 科学管理研究, 2012, **30**(6): 29-32.

[8] 吴菲菲,封红丽,黄鲁成. **基于震级法的新兴技术经济效应评估框架研究** [J].科学学与科学技术管理, 2012, **33**(3):94-101.

[9] 卢文光. **新兴技术产业化潜力评价及其成长性研究** [D]. 北京：北京工业大学, 2008.

[10] 魏国平. **新兴技术管理策略研究** [D]. 浙江：浙江大学, 2006.

[11] 王鹏. **战略性新兴技术辨识方法研究** [D]. 辽宁：大连大学, 2013.

[12] Chang P T,Huang L C,Lin H J. **The fuzzy Delphi method via fuzzy statistics and membership function fitting and an application to the human resources** [J]. Fuzzy Sets & Systems, 2000,**112**(3):511-520.

[13] Ishikawa A,Amagasa M,Shiga T,et al. **The max-min Delphi method and fuzzy Delphi method via fuzzy integration** [J]. Fuzzy Sets & Systems, 1993, **55**(3):241-253.

[14] Tran T A,Daim T. **A taxonomic review of methods and tools applied in technology assessment** [C].Management of Engineering & Technology, Portland International Center for. 2007:1651-1660.

[15] Shen Y C,Chang S H,Lin G T R,et al. **A hybrid selection model for emerging technology** [J]. Technological Forecasting & Social Change, 2010, **77**(1):151-166.

[16] Bart Verspagen. **Measuring Intersectoral Technology Spillovers: Estimates from the European and US Patent Office Databases** [J]. Economic Systems Research, 1997, **9**(1):47-65.

[17] Kostoff R N,Schaller R R. **Science and technology roadmaps** [J]. IEEE Transactions on Engineering Management, 2001, **48**(2):132-143.

[18] Verspagen B,A.B.JaffeM. **TrajtenbergPatents,Citations and Innovations: A Window on the Knowledge Economy2002MIT Press** [J]. Research Policy,2004, **33**(10):1709 – 1711.

[19] 陈亮,张志强. **一种基于专利文本的技术系统构成识别方法** [J]. 图书情报工作, 2014,**58**(10):134-137.

[20] 李倩. **基于专利的新兴技术弱信号识别方法研究** [D].北京：北京工业大学, 2014.

[21] 王凌燕,方曙,季培培. **利用专利文献识别新兴技术主题的技术框架研究** [J].图书情报工作, 2011, **55**(18):74-78.

[22] 刘同,真溱,汤珊红. **利用专利分类,识别和分析新兴技术**[J].情报理论与实践,2015(6):145-145.

[23] Jun S,Lee S J. **Emerging Technology Forecasting Using New Patent Information Analysis** [J]. International Journal of Software Engineering & Its Applications, 2012, **6**(3).

[24] Chang P L,Wu C C,Leu H J. **Using patent analyses to monitor the technological trends in an emerging field of technology: a case of carbon nanotube field emission display** [J]. Scientometrics, 2010, **82**(1):5-19.

[25] Yoon B,Park Y. **Development of New Technology Forecasting Algorithm: Hybrid Approach for Morphology Analysis and Conjoint Analysis of Patent Information** [J]. IEEE Transactions on Engineering Management, 2007, **54**(3):588-599.

[26] Bengisu M,Nekhili R. **Forecasting emerging technologies with the aid of science and technology databases**[J]. Technological Forecasting & Social Change, 2006, **73**(7):835-844.

[27] Daim T U,Rueda G,Martin H,et al. **Forecasting emerging technologies: Use of bibliometrics and patent analysis** [J]. Technological Forecasting & Social Change, 2006, **73**(8):981-1012.

[28] Griliches Z. **Patent statistics as economic indicators: a survey** [R]. National Bureau of Economic Research,1990.

[29] Trappey A J C,Trappey C V,Wu C Y,et al.**A patent quality analysis for innovative Technology and product development** [J]. Advanced Engineering Informatics, 2012, **26**(1): 26-34.

[30] Leydesdorff L,Kushnir D,Rafols I. **Interactive overlay maps for US patent(USPTO)data based on International Patent Classification (IPC)**[J]. Scientometrics,2014, **98**(3): 1583-1599.

[31] Criscuolo P. **The'home advantage'effect and patent families.A comparison of OECD triadic patents,the USPTO and the EPO** [J]. Scientometrics, 2006, **66**(1): 23-41.

[32] 李睿,张玲玲,郭世月. **专利同被引聚类与专利引用耦合聚类的对比分析** [J]. 图书情报工作,2012,08:91-95.

[33] Pawlowski S,Holman T. **Cluster controller for memory and data cache in a multiple cluster processing system: U.S. Patent 6,151,663[P].** 2000-11-21.

[34] Erdi P,Makovi K,Somogyvári Z,et al. **Prediction of emerging technologies based on analysis of the US patent citation network** [J]. Scientometrics, 2013, **95**(1): 225-242.

[35] Cho T S,Shih H Y. **Patent citation network analysis of core and emerging technologies in Taiwan: 1997–2008** [J]. Scientometrics, 2011, **89**(3): 795-811.

[36] 刘倩楠. **基于专利引文网络的技术演进路径识别研究** [D]. 辽宁: 大连理工大学, 2010.

[37] 王贤文,徐申萌,彭恋,等. **基于专利共类分析的技术网络结构研究:1971—2010** [J]. 情报学报, 2013, **32**(2):198-205.

[38] Kim Y G,Suh J H,Sang C P. **Visualization of patent analysis for emerging technology** [J]. Expert Systems with Applications, 2008, **34**(3):1804-1812.

[39] 黄鲁成,卢文光. **基于属性综合评价系统的新兴技术识别研究** [J]. 科研管理, 2009,**30**(04):190-194.

[40] 李欣,王静静,杨梓,等. **基于SAO结构语义分析的新兴技术识别研究** [J]. 情报杂志,2016, **35**(03):80-84.

[41] Yoon B,Park Y. **A text-mining-based patent network: Analytical tool for high-technology trend** [J].Journal of High Technology Management Research, 2004, **15**(1):37-50.

[42] 乔林,张雄伟,史海宁,等.**深度学习应用中的常见模型**[J].军事通信技术,2016(1).

[43] Deng L,Li J,Huang J T,et al. **Recent advances in deep learning for speech research at Microsoft** [C]. 2013:8604-8608.

[44] 陈硕.**深度学习神经网络在语音识别中的应用研究** [D].广东：华南理工大学, 2013.

[45] 林妙真.**基于深度学习的人脸识别研究** [D].辽宁：大连理工大学, 2013.

[46] He K,Zhang X,Ren S,et al.**Deep Residual Learning for Image Recognition** [J].Computer Science, 2015.

[47] Baccouche M,Mamalet F,Wolf C,et al.**Sequential Deep Learning for Human Action Recognition** [C]. International Conference on Human Behavior Unterstanding. Springer-Verlag,2011:29-39.

[48] Xie L,Pan W,Tang C,et al. **A pyramidal deep learning architecture for human action recognition** [J]. International Journal of Modelling Identification & Control, 2014, **21**(2):139-146.

[49] Akbar S,Khan M N A. **Critical Analysis of Density-based Spatial Clustering of Applications with Noise (DBSCAN) Techniques** [J]. International Journal of Database Theory & Application,2014, 7.

[50] Pollard D. **Quantization and the method of k-means** [J]. IEEE Transactions on Information Theory, 1982, **28**(2):199-205.

[51] Sperandei S. **Understanding logistic regression analysis** [J]. Biochemia Medica, 2014, **24**(1):8-12.

[52] Flach P A,Lachiche N. **Naive Bayesian Classification of Structured Data** [J]. Machine Learning, 2004, **57**(3):233-269.

[53] Neumann J,Schnorr C,Steidl G. **Combined SVM-Based Feature Selection and Classification** [J]. Machine Learning, 2005, **61**(1-3):129-150.

[54] 黄志华. **基于BP算法的多层感知器网络原理及程序实现**[J]. 嘉应学院学报, 2015,**33**(8):18-21.

[55] Qin S,Zhang B,Wang W,et al. **Throat Polyp Detection Based on the Neural Network Classification Algorithm** [M]. The Proceedings of the Third International Conference on Communications, Signal Processing, and Systems. 2015:847-855.

[56] Schmidhuber J. **Deep learning in neural networks: An overview** [J]. Neural Networks, 2014,**61**:85-117.

[57] Furukawa, T.,Mori,K.,Arino,K.,Hayashi,K.,Shirakawa, N. **Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions** [J]. Technological Forecasting & Social Change, 2014,**91**(0), 280 – 294.

[58] **Classes in the U.S. Patent Classification** [EB/OL]. `http://www.uspto.gov/page/classes-us-patent-classification-system-dates-established`.

致 谢

在完成研究生学位论文之时，回首研究生走过的岁月，三年以来自己点点滴滴的进步与成长无不凝聚着周围同伴和亲人的心血与关心。

在研究生学习阶段，有睿智博学而又坦诚的导师程戈副教授对我循循善诱，博之以文，约之以礼，使我欲罢不能。此外，程老师严谨的治学态度和高尚的师德在学习和生活以及科研等方面带来了无限的正能量，这种能量会让我终生受益。师生之情更是我人生中最宝贵的一笔财富。在此，特向程戈老师致以最崇高的敬意和最真诚的感激!

此外，由衷的感谢Moses博士、张振宇、张云、李咏林师弟、刘奎师弟、何晶晶师妹、张智师弟、陈波师弟、周金海师弟、李亚君师妹、钟杰师妹、章盼师弟等同门在学习、科研和生活上对我的帮助与关心。在每次讨论课上，各位积极思考问题的态度、敏锐的洞察力以及激烈的讨论都给我带来了莫大的帮助和前进的动力。

感谢父母、家人以及身边的朋友们对学业上的支持和生活上的资助，你们是我人生中最重要的精神支柱。

衷心的感谢数学与计算科学学院的所有老师，谢谢你们给我提供了这么好的学习和科研环境。

最后，特别感谢从百忙之中抽空对论文进行批阅与指正的评审专家们和答辩老师们。

攻读硕士学位期间的主要研究成果

一、发表的学术论文

1. Ge Cheng,Kyebambe Moses Ntanda,Yunqing Huang,Chunhui He,Zhenyu Zhang.Improving Patent Classification by Learning Weights through Back-propagation [J].Intelligent Automation and Soft Computing.2017.(修改再投)
2. Ge Cheng,Kyebambe Moses Ntanda,Yunqing Huang,Chunhui He,Zhenyu Zhang.Forecasting Emerging Technologies: A Supervised Learning approach through Patent Analysis [J].Technological Forecasting & Social Change.2017.(修改再投)

二、申请的发明专利

1. 程戈, 张振宇, 李强, 李聪, 张云, 何春辉. 基于最大熵模型的团体比赛结果预测方法: 中国. 201510174490.1 [P]. 2015-04-14.
2. 程戈, 张振宇, 李强, 李聪, 张云, 何春辉. 基于pagerank算法的体育竞赛团体实力预测方法: 中国. 201510174275.1 [P]. 2015-04-14.

三、参与的科研项目

1. 国家自然科学基金青年项目《基于轻量虚拟化可信基的可信计算环境构建机制研究》, 项目编号: 61202397.

四、个人简历

1. 2010-09至2014-06 , 在湖南城市学院应用数学专业学习并获得理学学士学位.
2. 2014-09至2017-06 , 在湘潭大学攻读硕士，研究方向为信息处理及应用软件.