1. 绪论

企业信用评级是指信用评级机构利用某种方法或者模型对制造业企业、服务业企业、建筑业企业的信用进行评级。随着评级方法的不断完善，企业信用评级方法考察的范围越来越广：服务能力、发展潜力、技术能力、战略眼光等。

目前投资者与企业之间存在严重的信息不对称的情况，即企业内部对自己有深刻的认知，然而投资者缺少企业的相关信息，或者需要从海量的数据中寻找企业的相关信息，这种情况不仅为投资者投资企业增加了难度，并且增加了投资者自身的收益风险。

当下社会处于经济转型期，很多经营者与投资者依然存在“旧思想”和“旧观念”，即国家政府掌控企业的资金。另一方面，我国的企业评级起步较晚，没有受到广泛的普及。然而，企业信用普及是一种最直观有效的信息，为投资者投资企业提供了重要的参考。传统的企业信用评级方法利用的信息之间往往存在复杂的非线性关系，因此传统的企业评级技术不能很好的适应当前的企业信用系统。

近年来，随着企业评级技术与计算机硬件与软件的不断发展，企业各方面的数据呈指数增长，使得企业评级的准确度与参考价值得到了明显地提高。然而，数据量的增加，也为企业信用评级带来了一些列的问题，具体来说：（1）企业的信用度与评价指标之间呈现出非线性关系。传统的企业评级的线性模型不再适用于当前的企业信用评级系统，需要一种能够反映属性之间复杂关系的模型；（2）参考指标多样化。传统的企业评级的综合模型在此情况下，即便是领域专家也不可能了解每一个属性，并且人工的方式易造成不公平的现象。因此，提出可靠高效的企业信用评级方法是一个亟待解决的重要问题。

神经网络模型在当前时代背景下得到了广泛的发展，例如，运用神经网络在高分辨率遥感图像的识别任务上，世界上最先进的高分辨率遥感图像的分辨率达到了分米级，合成孔径雷达（Synthetic Aperture Radar, SAR）已达米级，采用干涉测量技术的合成孔径雷达（Interferometric Synthetic Aperture Radar, InSAR）相对位移精度达到了毫米级，高光谱图像（Hyperspectral Image）的分辨率甚至达到了纳米级。另一方面，卫星重访周期的单位已缩短至天。上述特征使得拍摄的高分辨率遥感图像不仅具备数量多、多源化、异构化的特征，还具备高维度、多尺度、非平稳的内部特性，具备大数据的特点[1]。然而，适配于中低分辨率的传统遥感影像解译方法，在识别包含车辆、飞机、轮船等复杂特征的高分辨率遥感影像时效率低下，甚至无法准确识别。面对新时代高分辨率遥感影像目标识别的更高要求，亟需探索全新的、更加高效的高分辨率遥感影像目标识别方法与模型，用来提高高分辨率遥感图像的目标识别精度、识别技术的智能化水平[2-4]。很多工作表明神经网络在很多领域表现突出甚至达到了人类的水平。

* 1. 研究背景与意义

自改革开放以来，我国经济迅猛发展，企业已经成为推动我国经济发展的重要力量。然而随着时代的进步，企业的各个方面出现了一系列的问题例如：技术革新、发展战略、人才管理和企业信用等。其中企业信用评估这一领域在我国仍处于初级阶段，导致我国众多企业信用缺失、企业内部信用管理缺失、信用理念不强、企业对自身信用的利用不够充分。另外，信用的缺失不仅导致投资者对企业的信用度下降还使得消费者对企业产品持怀疑的态度，影响企业的长远发展。

随着时代的发展，我国逐渐重视企业和个人的信用度。自1987年，国家建立第一个企业信用评级机构，我国企业信用评级的观念渐渐深入人心，并且使得企业以及个人重视维护建立自身的信用度。进入新世纪以来，我国企业的信用评级机构、制度和技术得到不断的发展。新的时代带来新的问题，促使研究者将目光引入这一新兴领域，在已有方法的基础上，根据我国特色社会主义的发展模式，建立使用我国企业的信用评级方法。

目前我们正在进入一个智能化时代（互联网+及工业4.0时代），大数据、云计算和人工智能已经成为这个时代进步的三驾马车(（图1-1所示）)，它们分别为智能化时代提供数据、算力和算法层面的支持，从而成为各行各业技术革新和社会发展的重要引擎。

在智能化时代的背景下，云计算为企业信用相关的数据提供了大量的存储空间，使得海量、多维度的企业数据存储成为可能。研究者可以从大量存储的数据中识别、提取感兴趣的信息，为企业自身发展、企业规划与扩建、企业信用评级机构审查、投资、甚至社会发展的等领域带来巨大的益处，具体来说：（1）在企业自身发展方面：从企业自身的历史数据中提取企业发展的趋势，现在企业面临的挑战，查漏补缺；（2）在企业规划与扩建方面：通过获取大量企业自身的数据

图 1-1 拉动智能时代的三驾马车



Figure 1-1 Pulling the Troika of the Intelligent Age

以及其它企业的数据，获得当前的热门领域，根据自身的状况进行及时转型；（3）在企业规避风险方面：从企业自身的历史数据中提取发展趋势，从而得知自身的发展瓶颈以及个能遭遇的挑战，及时做好防备工作，面对挑战；（4）在评级机构方面：评级机构可以从大量的数据中综合地、全面地、多角度地对企业信用进行评级，充分利用先进的算法的性能，达到真实客观地评价企业的信用。

基于上述分析，遥感图像识别是运用遥感技术解决各种问题的关键步骤。高分辨遥感图像中的目标具有多样性与复杂性，具体来说：（1）在多样性方面：搭载在航空、航天等设备上的仪器拍摄的遥感图像由于拍摄角度的原因，图像中物体呈现出事物种类多、事物之间联系复杂，既包含要识别的目标，又包含诸多干扰物体，如山川、河流、各种建筑物等；（2）在复杂性方面：目标与目标之间、目标与非目标之间往往存在复杂的联系，且呈现出相互影响的特征。该分辨率遥感图像的这些特点为识别图像中的目标带来了挑战。传统的遥感图像识别技术（如：数字图像处理软件、统计模式识别方法等）存在严重依赖人工解译，解译精度较低等诸多问题。因此，如何高效、准确地识别高分辨率遥感图像中的目标是一个亟待解决的关键问题。

神经网络在文本处理、目标分类、图像识别、目标识别领域具有优秀的表现[5]。另一方面，随机计算机技术以及企业评级指标的不断发展，企业评级可利用的资源越来越多：数据多、模型多等，这使得神经网络运用到企业信用评级中成为可能。与传统方法相比，神经网络在基于文本挖掘后的数据基础上，可以识别属性间的非线性关系，由于模型是在数据中学习企业分类的逻辑，因此在分类的过程中避免了人员的参与，不仅在理论上提高了企业信用的分类精度还保证了企业信用评级的公平性。然而，如何从大量的数据中获取有效的信息，识别冗余信息，提取关键信息，去除冗余属性是一件具有挑战的任务。如何基于文本挖掘技术正确地将神经网络运用到企业信用评级中，自动、高效地获得企业信用等级具有重要的研究价值。

* 1. 研究内容与成果

企业信用评级问题实际上是企业信用的分类问题，在处理分类问题方面，神经网络具有优秀的特征表示能力、强大的自我学习能力和准确的目标识别能力。随着企业信息量不断的增大与学习模型的不断革新，利用神经网络对企业信用的评级也更加可靠。本文主要研究基于文本挖掘的神经网络在企业信用评级任务中的有效性与准确性，具体表现在以下2个方面：

1. **提出了基于文本挖掘的adahoost和神经网络企业评级框架(MBP-CCS)**：**MBP-CCN**利用文本挖掘相关的技术，得到相应的数据，训练自适应神经网络（SOM），然后利用adahoost算法得到强分类器。
2. **采用经验研究验证所提模型对企业信用评级分类的准确性以及有效性**：利用中国上市公司网站的信息，随机选取200个正常营业的制造业公司和50个非正常营业的制造业公司来训练模型，并利用测试用例集评估所提模型的正确性与准确性。
   1. 论文组织结构

本文的组织结构如下：

1. 绪论，主要包括课题的研究背景与意义、研究内容与成果以及论文的组织结构。
2. 背景介绍，主要包括神经网络和词袋模型的基础概念与理论，以及目标识别、卷积神经网络的国内外研究现状。
3. 深入讨论本文提出的基于文本挖掘的神经网络企业信用评级方法的框架。
4. 在开源的数据集上验证所提框架的有效性与准确性，主要包括研究问题阐述、实验对象说明、实验流程。
5. 通过图、表的方式回答所提的问题，并深入分析实验结果。
6. 研究工作的总结以及未来的展望。
7. 背景介绍

本章介绍卷积神经网络和词袋模型的相関概念、基础理论以及国内外研究现状。

* 1. 相关技术与概念

本节首先介绍卷积神经网络的基本特征，然后介绍几种经典的神经网络算法，最后介绍一种传统的遥感图像目标识别方法—词袋模型。



图2-1 LeNet-5的网络结构

Figure 2-1

* + 1. 卷积神经网络

现在的神经网络最早是由Lecun等人提出的7层网络结构（如图2-1）LeNet-5[7]，主要包括卷积层（C）、池化层（S）、全连接层（F）。通常卷积层紧跟在输入层之后，池化层在卷积层之后。神经网络通过重复输入层、卷积层与池化层这种连接结构增加神经网络的深度。为了将提取的特征映射为实际的输出，全连接层一般出现在神经网络的末端。接下来，详细介绍神经网络中的主要组件：

1. **输入层：**在训练和测试卷积神经网络的过程中，输入层的神经元数目通常是

固定的，负责接收大量的多维数组（数据）。与传统的全连接神经网络（如自适应神经网络[11]）相比,卷积神经网络不需要对数据进行归一化处理（即不需要限制数组中每一个维度的值），也不需要对数据进行中心化处理（即数据每一个维度的值减去平均值）[12]。

1. **卷积层：**在泛函分析中，卷积是通过两个已知函数和生成新函数的算

子，记为公式2.1。在卷积神经网络中，卷积运算起到了滤波的作用，将输入图像的局部区域与权值进行加权运算，其权值由一个的函数定义。卷积的具体表现

**(2.1)**

形式为公式2.2，其中表示图中坐标为的灰度值,为卷积核。公式2.1的运算结果为卷积核对相关输入的响应值，即为输出图像中的一个像素。

**(2.2)**

卷积神经网络区别于传统的全连接网络的重要特征是存在卷积层。该层首先利用卷积核对上一层的局部输出数据进行卷积运算，然后利用非线性激活函数将卷积的结果限定在某一个范围内，使得模型具备非线性的特征。通常，利用公式2.3和2.4进行卷积运算，其中表示卷积运算的输出值，是卷积层的输入，表

**(2.3)**

**(2.4)**

示卷积核，表示卷积的加性偏项，是该卷积层的输出特征图，表示一个激活函数。与传统全连接的权重更新方式类似，在训练模型的过程中卷积神经网络的卷积核通过反向误差传播算法学习得到，并且每一个卷积层包含多个卷积核。不同的卷积核与输入的图像进行卷积运算，得到输入图像的多样化特征。一般情况下，激活函数是非线性函数（常用的激活函数如表2-1，每一种激活函数对应的函数图像如图2-2），根据输入的值计算出相应的输出并将输出的值控制在一定的范围内，使得卷积神经网络可以捕获图像的非线性特征。表2-1展示的激活函数中，Sigmoid函数是全连接神经网络以及传统卷积神经网络中常用的激活函数。图2-2(a)表明Sigmoid函数是单挑递增的光滑函数，函数值的范围限制在。然而，Sigmoid函数在数据正向传播时计

表2-1 常用的激活函数

Table 2-1

|  |  |  |  |
| --- | --- | --- | --- |
| 激活函数 | 定义 | 参数 | 图像 |
| Sigmoid |  | -- | 图2-2(a) |
| Tanh |  | -- | 图2-2(b) |
| ReLU |  | -- | 图2-2(c) |
| LeakyReLU |  |  | 图2-2(d) |
| PReLU |  | 是学得的参数 | 图2-2(d) |
| RReLU |  |  | 图2-2(d) |
| ELU |  | 是预先设参数 | 图2-2(e) |

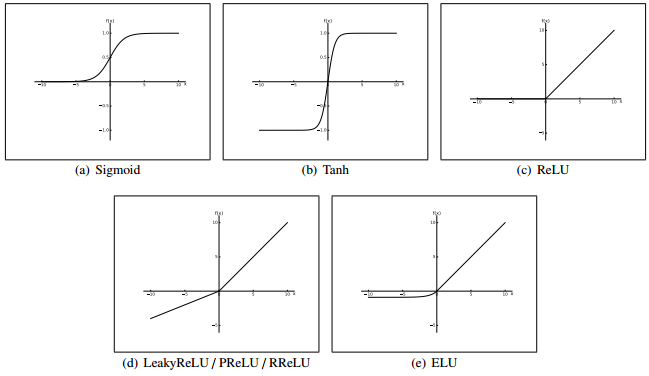


图2-2 激活函数对应的图像

Figure2-2

算量大，在误差反向传播时容易出现梯度消失的情况，在深层的神经网络中表现欠佳。随着卷积神经网络架构的不断改进以及算法的不断更新，卷积神经网络的层次不断增加（AlexNet有8层，VGGNet有19层，ResNet最大有152层，GooLeNet有22层），ReLU激活函数逐步代替Sigmoid函数，得到广泛的使用。与Sigmoid函数相比，ReLU激活函数有效地缓解了误差反向传播时梯度消失的问题，直接以监督的模式训练神经网络，并且具有收敛速度快的特点。然而，当一个较大的梯度流经神经元时，传统的ReLU激活函数容易使得该神经元不能被激活，造成模型训练不充分的问题。为此，LeakyReLU、PReLU等改进的ReLU激活函数陆续被提出，并取得良好的效果。

1. **池化层：**该层将池化策略通到卷积层输出的图像特征，降低特征图的大小与分辨率，有效地减少了参数的卷积神经网络中参数的个数，也具有防止过拟合的作用。典型的采样方法有两种：（1）对特征图像相邻像素中取最大值；（2）计算相邻像素的平均值。图2-3展示了选取最大池化方法作用在图像后的结果，假设池化的窗口为。尽管最大和平均的池化方法在卷积神经网络中表现良好，一些

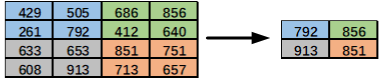


图2-3 最大池化法作用实例

Figure2-3

研究者提出了更好的池化方法：He等人提出了空间金字塔池化方法[13]，对于任意大小的特征图首先分成16、4、1个块，然后在每个块上运用最大池化方法，最后将池化后的特征拼接得到一个固定维度的输出上。该方法的主要有点是对于任意尺寸的输入产生固定大小的输出。

1. **全连接层和输出层：**卷积神经网络的末端是全连接层和输出层。输入的图

像经过若干次卷积运算与池化后，得到高水平的抽象特征，然后每一个神经元将提取的图像特征均输入到全连接层（即该层与上一层的连接结构具有全连接的特征），最后通过输出层的神经元输出模型最后结果。这种连接方式不仅有效地降低了卷积神经网络中参数的个数还充分地提取图像复杂、多样化的特征，还能挖掘图片更深层次的信息。输出层是卷积神经网络的最后一层，该层的神经元数目在训练和测试的过程中具有不变性，主要功能是对图像做出最终的判断并输出。

* + 1. 卷积神经网络的特点

从卷积神经网络的提出到第一次成功的运用在图像识别上，该技术一方面收益于计算机硬件与软件的发展，另一方面卷积神经网络自身也具有成功的特点。在图像分类与图像的目标识别领域，卷积神经网络在工作的过程中不需要人员的参与，自动地从图像中提取特征，从数据中获得识别图像目标或者对图像进行分类的逻辑，极大地解放了人员的参与；另一方面，卷积神经网络内部层与层之间不同层内部的机制，也使得卷积神经网络的成功成为必然，例如卷积神经网络通过局部感受野、权值共享和降采样的方式使得图像做位移、缩放等变换时具有不变性，具体来说：

1. **局部感知野：**该特征受到生物局部神经元获取视野方式的启发，研究者认为

只需要对输入图像的局部进行感知，即某一输出层一个点的值对应输入层某一区域，随着卷积神经网络层数不断加深，高层神经元可以获得高水平的原始图像的全局抽象特征。与全连接的神经网络相比，在保证提取特征效果不变的条件下卷积神经网络中参数数目更少。假设输入图像的像素为，全连接神经网络的隐藏层和卷积神经网络的卷积层均具有100个神经元且每一个神经元与的局部图像相连，则对于全连接网络，参数的数目为：；然卷积神经网络的参数数目为：（参数的数目相对于全连接神经网络减小了100倍）。通过上述分析可知，卷积神经网络极大地化简了神经网络，提高了神经网络的训练与测试时间，节省了成本。

1. **权值共享：**随着问题要解决问题的复杂程度不断增大，卷积神经网络的深度

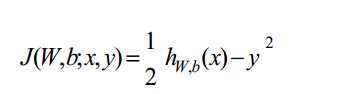
呈现出加深的趋势，这种架构使得卷积神经网络的参数不断增多。另一方面对了提高模型的准确度，训练数据的个数是海量的，势必需要更多的时间进行训练与测试。基于上述问题，研究者探究进一步减小卷积神经网络参数的方法，提出了权值共享的方法。这种想法基于生物的神经网络和图像不同部位具有一定的相似性这一观察。该想法具体的实现是同一个卷积核在与图像的不同部位进行卷积运算时的参数保持不变，从而在卷积层的感受野的基础上进一步减小参数的数目。在上述例子中，卷积层每一个神经元与的局部图像相连，即参数的数目是100，由于这1000个神经元共用这个100个参数，因此一个卷积核的参数数目为100；否则参数数目为。一般情况下，为了充分提取图像的特征信息，卷积层往往具有多个卷积核，然而多个卷积核需要的参数数目与卷积核的个数呈现线性关系，避免出现参数数目爆炸的情况。

1. **降采样：**局部感受野和权值共享有效地减少了卷积神经网络中的参数数目，

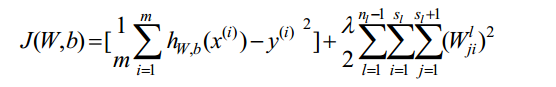
极大地降低了模型的训练与测试时间。但是当输入图像比较大时，仍会降低模型的训练与测试速度。基于上述例子，输入的图像经过第一个卷积层之后得到个特征图，每一个特征图与输入的图像进行卷积得到维的卷积特征，一共存在8100个特征图，总计维的特征向量。如此庞大的维数不仅需要大量的时间，还容易出现过拟合的现象。为了缓解这一问题，研究者提出了降采样的方法，即将特征图映射成一个小尺寸的特征图，但是扔保留了特征图中的信息，相当于对特征图进行了抽象。由此，在提取输入图像特征信息的基础上不仅降低了卷积神经网络的参数数目，还降低了输入的维数。

1. **误差反向传播：**传统的全连接神经网络由于存在阈值函数，使得基于误差调

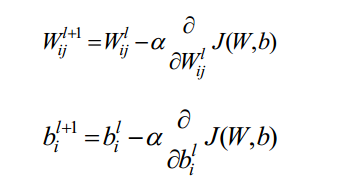
整权重是一个难以解决的问题。直到梯度下降法的提出，缓解了全连接神经网络的权重更新问题。该算法需要数据标签，并根据模型的输出值与标签进行对比，得到隐藏层与输出层每一个神经元的误差。由于输出层每一个神经元的输出误差是由隐藏层中每一个神经元的误差与权重加权的和导致的，因此输出层所有神经元的误差与隐藏层某一个神经元的权重加权并求和，得到隐藏层的误差。重复以上步骤直到更新第一个隐藏层与输入层之间的权重。研究者基于全连接神经网络更新误差的方式，提出了批量梯度下降法，并成功运用到卷积神经网络中：假设存在一个包含个样本的训练集 。从中选取样本并输入模型后，运用梯度下降法对其求解，神经网络的损失函数为：



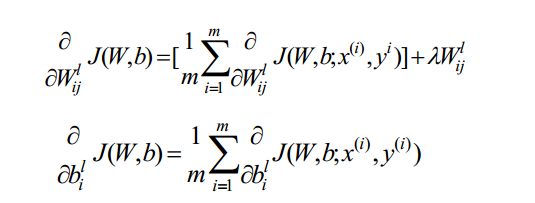
则包含个样本的数据集的代价函数可以表示为



从上式可以看出，代价函数是凹函数，即具有最小值，但存在多个极小值。为了避免结果出现在错误的极小值位置，可以多次输入来进行调整。在实际的问题解决过程中，梯度下降法的表现良好。通过判定目标函数是否收敛来判断模型是否训练结束。通过求导，可以得出每一次输入数据之后参数的变化情况，并通过以下公式进行更新：



其中，\alph是学习速率。对于整个数据集，代价函数的偏导如下：



综上所述，卷积神经网络具有强大能力的同时，其内部层次之间以及层次之中的特点使得卷积神经网络的参数极大地降低，输入数据的维度得到进一步降低，极大地节省了测试与训练的时间。使得卷积神经网络在实际的应用中表现良好，甚至达到了人类的水平。针对某一种具体类型的问题，或者在LeNet-5模型的基础上，研究者们先后提出了AlexNet[5]、GooLeNet[8]、VGG-16[9]、ResNet[10]等被广泛学习的神经网络模型，表2-2总结了常见神经网络最重要的特征，本文涉及的神经网络模型详细地描述在2.1.3章节。

* + 1. 神经网络框架简介

深度学习技术在各个领域取得了令人惊叹的成绩，例如特斯拉、谷歌以及百度的无人驾驶汽车；人脸识别；AlphGo战胜了世界排名第一的柯洁等，使得深度学习收到了广泛的关注。为了方便更多的人从事深度学习相关的研究，研究者们基于不同的高级语言（例如Java、python、C++等）开发出了一系列的框架，例如TensorFlow、Caffe、Keras等。Clark[14]等人为了研究开源的深度学习框架受欢迎的情况，基于框架的关注度以及提交代码的数量，对16个常用框架流行程度进行了排序，结果展示在图2-3。从图中可以明显看出TensorFlow[15]的关注度最高，说明该框架是目前使用最多的深度学习框架。Dlib框架的关注度最低，说明该框架使用的人数最少。TensorFlow框架不仅可以用来研究深度学习相关的问题，还可以用来做其它方面的数值研究，该框架具有可移植性，即用户可以在

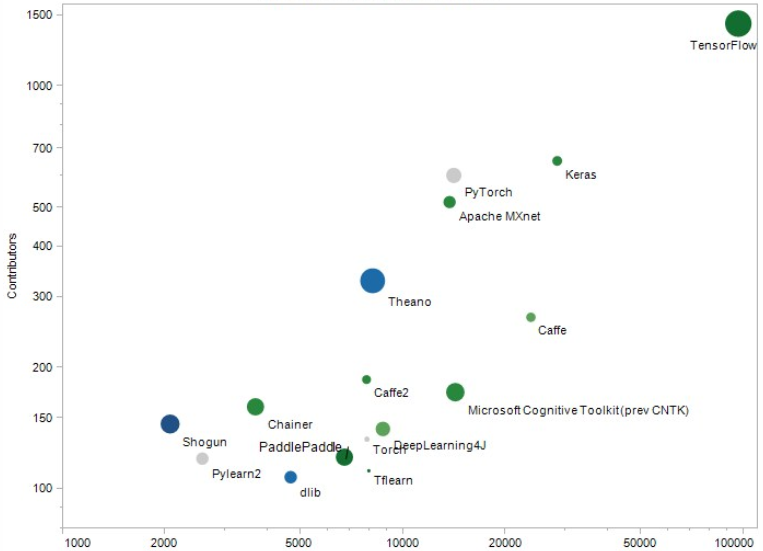


图 2-3 16种深度学习开源框架关注度排名

Figure2-3

多个平台使用该框架。例如台式计算机中的一个或多个CPU（或GPU），服务器，移动设备等。TensorFlow最初是由Google公司开发出来的，由于其简介的代码、友好的接口以及可移植性，使得该框架收到广泛的关注。后来，发布到开源社区，越来越多的研究者改进架构、修复故障等，使得该框架具有非常高的可用性。该框架的主要特征如下：

1. **高度的灵活性：**TensorFlow不仅可以用来进行神经元网络的研究还可以将研究者的计算转化为为一个数据流图。该框架提供了有用的工具来帮助研究者组装“子图”（常用于神经网络）。此外，研究者可以自己在该框架的基础上定制自己的“上层库”。
2. **真正的可移植性：**很多开源框架（如caffe）只能在含有CPU和GPU的平台

上运行，例如台式机、服务器、手机移动设备等。Tensorflow可以在没有特殊硬件的前提下，调用接口，实现既定的功能。

1. **自动求微分：**基于梯度的机器学习算法会受益于Tensorflow具有自动求微分

的能力，极大地便利了基于梯度的机器学习算法。研究者只需要定义预测模型的结构，将该结构和目标函数结合在一起，然后添加数据，该框架将自动为研究者计算相关的微分导数。

1. **多语言支持：**为了适应开发者技能具有多元化的特征，该框架不仅提供了多

种高级语言（C++、python等）的接口（实现了某种功能的函数）而且提供了用户交互的界面。研究者在用户交互界面可以使用自己掌握的语言来构建与执行数据。

1. **性能最优化：**该框架给予了线程、队列、异步操作等最佳的支持，可以将研

究者可用的硬件计算潜能全部发挥出来。该框架还支持自动地将图像中的计算元素分配到不同设备上，通过这种方式来实现充分利用计算资源的目的。

* + 1. AlexNet卷积神经网络

AlexNet卷积神经网络[5]是由Alex Krizhevsky等人在2010年图像分类竞赛中提出的，该神经网络成功地将120W张高分辨率的图像分为1000个类别，并取得了冠军。该框架需要两个GPU分别跑前五个卷积层，从第一个到第三个全连接层共享参数，每个GPU分摊二分之一的参数，最后将生成的特征图分配给GPU进行计算，具体结构展示在图2-4。

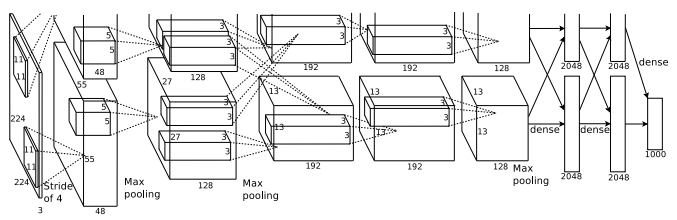


图2-4 AlexNet卷积神经网络架构

Figure 2-4 The Architecture of AlexNet

AlexNet神经网络具有以下特征：

1. 收敛/训练速度加快，用了非饱和神经元（ReLU，即线性矫正单元）。ReLU是

一种非线性神经元（当输入x大于0，则输出为x；输入x小于0，则输出0）较之前的激活函数计算速度更快，收敛也更快；

1. 模型并行。文中并没写是模型并行，但我看介绍认为是一种（有数据并行的）

模型并行。训练过程用了两个GTX580 GPU 3GB（一个gpu无法cover住），除了第三层卷积、最后的两个全连接层（两个GPU将全连接拆成各自两部分）和 Softmax（汇聚到单个GPU处理），其余部分处理可以看作两个GPU的数据并行（AlexNet架构图的上部分和下部分对应两个GPU各自的处理流程，单GPU的话，就是 group=2 这个可以看后文中通过netscope可视化出的网络结构）。相比单GPU没加快多少，同时某些层两个GPU共享参数，可以减少显存占用这样以便在一次参数更新放更多的图片加入训练。发现双卡比单卡top1和top5的error有下降（我认为是作者忘了改SGD学习率导致，先跑one-gpu net再跑two-gpu net）；

1. 正则化——数据扩增和dropout。在数据扩增方面，一般的做法是在已有数

据的基础上做某种变化，进而在变化的过程中找到某种规律。在保证数据多样性的同时，相似的图像输入到卷积神经网络中，这种方式使得卷积神经网络的权重不断更新，在输入到输出的过程中映射图像的本质；在dropout方面，模型在学习的过程中，缺少dropout模型的训练很快会出现过拟合的现象。

* + 1. GooLeNet卷积神经网络

GoogLeNet提出最直接提升深度神经网络的方法就是增加网络的尺寸，包括宽度和深度。深度也就是网络中的层数，宽度指每层中所用到的神经元的个数。但是这种简单直接的解决方式存在的两个重大的缺点。

(1) 网络尺寸的增加也意味着参数的增加，也就使得网络更加容易过拟合。

(2) 计算资源的增加。

因此想到将全连接的方式改为稀疏连接来解决这两个问题。由Provable bounds for learning some deep representations.提到数据集的概率分布由大又稀疏的深度神经网络表达时，网络拓扑结构可由逐层分析与输出高度相关的上一层的激活值和聚类神经元的相关统计信息来优化。但是这有非常多的限制条件。因此提出运用Hebbian原理，它可以使得上述想法在少量限制条件下就变得实际可行。通常全连接是为了更好的优化并行计算，而稀疏连接是为了打破对称来改善学习，传统常常利用卷积来利用空间域上的稀疏性，但卷积在网络的早期层中的与patches的连接也是稠密连接，因此考虑到能不能在滤波器层面上利用稀疏性，而不是神经元上。但是在非均匀稀疏数据结构上进行数值计算效率很低，并且查找和缓存未定义的开销很大，而且对计算的基础设施要求过高，因此考虑到将稀疏矩阵聚类成相对稠密子空间来倾向于对稀疏矩阵的计算优化。因此提出了VGG-16卷积神经网络。inception结构的主要思想在于卷积视觉网络中一个优化的局部稀疏结构怎么样能由一系列易获得的稠密子结构来近似和覆盖。上面提到网络拓扑结构是由逐层分析上一层的相关统计信息并聚集到一个高度相关的单元组中，这些簇（单元组）表达下一层的单元（神经元）并与之前的单元相连接，而靠近输入图像的底层相关的单元在一块局部区域聚集，这就意味着我们可以在一块单一区域上聚集簇来结尾，并且他们能在下一层由一层1x1的卷积层覆盖，也即利用更少的数量在更大空间扩散的簇可由更大patches上的卷积来覆盖，也将减少越来越大的区域上patches的数量。研究者在设计inception的结构时，限制了inception结构中滤波器的大小（），来达到避免patch对齐的问题。为了在高层能提取更抽象的特征，就要减少其空间聚集性，因此通过增加高层inception结构中的积数量，捕获更大面积的特征。

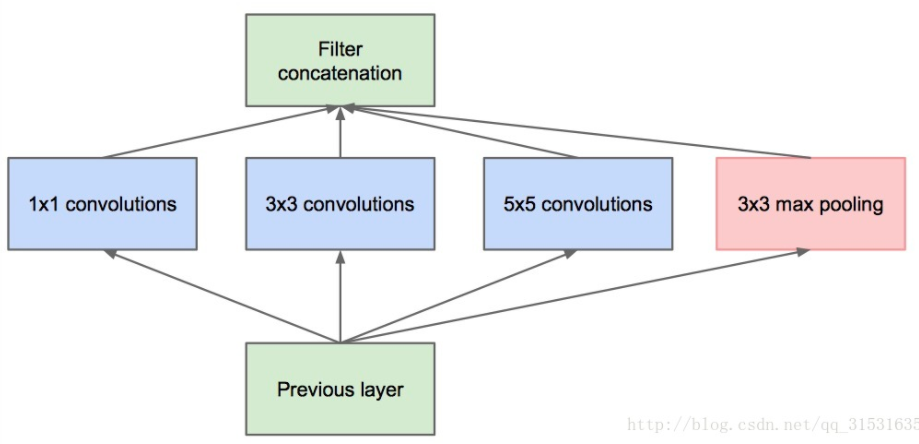


图2-5 inception 的架构

Figure 2-5 The Architecture of inception

* + 1. ResNet卷积神经网络

它差不多是当前应用最为广泛的CNN特征提取网络。它的提出始于2015年，作者中间有大名鼎鼎的三位人物He-Kaiming, Ren-Shaoqing, Sun-Jian。绝对是华人学者的骄傲啊。

VGG网络试着探寻了一下深度学习网络的深度究竟可以深几许以能持续地提高分类准确率。我们的一般印象当中，深度学习愈是深（复杂，参数多）愈是有着更强的表达能力。凭着这一基本准则CNN分类网络自Alexnet的7层发展到了VGG的16乃至19层，后来更有了Googlenet的22层。可后来我们发现深度CNN网络达到一定深度后再一味地增加层数并不能带来进一步地分类性能提高，反而会招致网络收敛变得更慢，test dataset的分类准确率也变得更差。排除数据集过小带来的模型过拟合等问题后，我们发现过深的网络仍然还会使分类准确度下降（相对于较浅些的网络而言

正是受制于此不清不楚的问题，VGG网络达到19层后再增加层数就开始导致分类性能的下降。而Resnet网络作者则想到了常规计算机视觉领域常用的residual representation的概念，并进一步将它应用在了CNN模型的构建当中，于是就有了基本的residual learning的block。它通过使用多个有参层来学习输入输出之间的残差表示，而非像一般CNN网络（如Alexnet/VGG等）那样使用有参层来直接尝试学习输入、输出之间的映射。实验表明使用一般意义上的有参层来直接学习残差比直接学习输入、输出间映射要容易得多（收敛速度更快），也有效得多（可通过使用更多的层来达到更高的分类精度）。

当下Resnet已经代替VGG成为一般计算机视觉领域问题中的基础特征提取网络。当下Facebook乍提出的可有效生成多尺度特征表达的FPN网络也可通过将Resnet作为其发挥能力的基础网络从而得到一张图片最优的CNN特征组合集合。

若将输入设为X，将某一有参网络层设为H，那么以X为输入的此层的输出将为H(X)。一般的CNN网络如Alexnet/VGG等会直接通过训练学习出参数函数H的表达，从而直接学习X->H(X)。

而残差学习则是致力于使用多个有参网络层来学习输入、输出之间的参差即即学习。其中这一部分为直接的识别映射，而则为有参网络层中输入输出之间的残差。

* + 1. Fast R-CNN卷积神经网络

Fast R-CNN借助多任务损失函数，实现了同时识别目标和位置修正两大功能，使得研究者不再需要对网络分布训练。Fast R-CNN独特的架构（如图2-6）使得研究者在训练该模型的过程中不需要大量的内存来存储中间数据，极大地降低了卷积神经网络对计算机硬件的硬性要求。图2-6中，卷积-池化层的功能与上文讲述的LeNet-5的功能类似，具有自动提取输入图像特征、降低卷积神经网络的参数数目以及较少输入数维数的作用。该结构识别的图像特征包括了要识别的目标特征。当前，大多数研究者利用选择搜索算法对自动识别输入图像的目标区域（TF）。一般情况下，TF包含两部分内容：（1）要识别的目标；（2）目标背景或与目标紧密的相关的其它物体。该算法提取的目标区域的大小具有多样性，然而该结构的全连接层对输入的特征向量的维数具有不变性的要求。因此，该模型将LeNet-5模型最后一个池化层替换为兴趣池化层，用来将不同位数的特征图统一化，即生成维数相同的特征向量，最后输入每一个候选区域特征的类型得分。

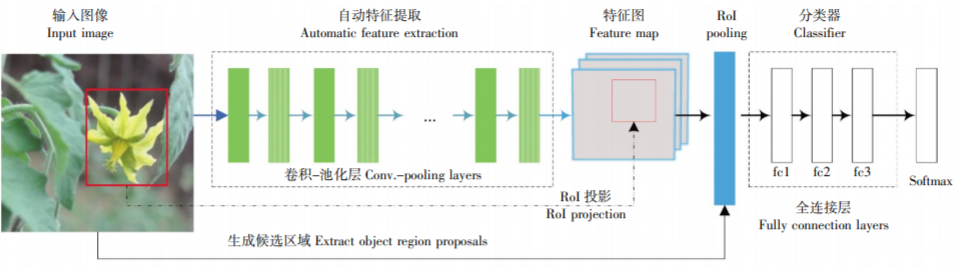


图2-6 Fast R-CNN卷积神经网络架构

Figure 2-6 The Architecture of Fast R-CNN

* + 1. 词袋模型

词袋模型最早是用来解决文档领域中的匹配问题，认为文档中每一个词都是独立出现的，也就是文档中词语与词语之间不具有关联性，将文档当成一个纯粹的填装词语的袋子，忽略了词语之间的顺序、语句的语法。该模型首先提取感兴趣的关键词，然后将文档表示成与顺序无关的关键词组合，最后利用统计的方法对文档中关键词出现的频率进行计算。该模型具有简单、易于理解的特点，研究者们将该模型成功地应用到分类任务中并取得了较好的结果[16,17]，主要包括提取和描述特征，创建字典，量化特征和训练分类器4个过程，具体来说：

1. **特征提取和描述：**一般情况下，输入图像的内容具有多元化与多样化的特征，

识别图像中的目标首先需要将目标中图像中区分并且提取出来。因此，图像特征提取和描述是图像分类以及图像目标识别任务中的第一步，该步骤主要任务是提取图像最核心的特征，删除无关元素，用提取的特征来表示输入的图像。最主要的图像特征提取与描述的传统技术是SIFT描述子。为了提高图像特征提取与描述的精度，本文首先利用稠密采样的方法，提取输入图像中的局部特征，然后再利用SIFT描述子提取图像的特征。

2) **视觉词典生成：**在上一步的基础上，得到了输入图像大量的特征，生成视觉词典的过程就是对特征域分区的过程，每一个分区的中心就是所谓的视觉单词。分区的方式有很多，常用的技术为K-means聚类。当对特征域分区之后，新的特征利用最近邻方法技术到每一个视觉单词的最近距离，并认为该视觉单词可以表达该特征。根据上述分析，可以得到视觉单词的数量与K-means的聚类中心数目相同。

3) **特征量化：**在量化特征的过程中，最近邻算法用来计算输入的特征图与所有视觉单词之间的距离的最小值，并得到可以表达该输入特征的视觉单词，然后利用统计学的知识，得到该视觉单词出现的数目，并绘制相应的直方图。传统的机器学习的分类算法中，支持向量机（SVM）是广受欢迎的分类器，在卷积神经网络出现前，该分类器是被研究最多的分类器。SVM希望找到一个超平面，并通过该超平面对不同的数据进行分类，其目标函数可以表示为：

C:\Users\daihe\AppData\Local\Temp\1551420043(1).png

约束条件为：

C:\Users\daihe\AppData\Local\Temp\1551420079(1).png

其中是超平面的法向量，为惩罚因子，为松弛向量。本文采用 SVM 进行分类，选用径向基核函数。

* 1. 国内外研究现状

本节介绍相关技术的研究现状，包括：目标识别和卷积神经网络。

* + 1. 目标识别的研究现状

传统的目标识别方法通常需要设计人员设计复杂的特征，需要借鉴大量的领域知识并经过巧妙设计后，在特定情况下取得了较好的效果，但这些人工设计特征无法充分体现图像中的具体含义与复杂的高层语义信息，使得特征的抽象和泛化能力较弱。

近年来，大数据为科学发展提供了大量、多元的数据，云计算技术为这些大量的数据提供了足够的存储空间，研究者利用不同的技术与原理来充分利用这些技术带来的益处，推动社会不断发展，经济不断提高。其中深度学习最近收到了人们的广泛关注，卷积神经网络在世界级各大竞赛以及实际的应用过程中表现良好，受到研究者与工业界的普遍认可。

在图像识别领域内，传统的技术需要大量的人工参与，例如人工设计或者提取研究对象的特征，模型参数调优等，极大地限制了图像分类以及图像目标识别技术的准确性。卷积神经网络的出现，为该领域带来了曙光，该模型不需要人工提取图像特征，并且参数的调整过程是完全自动化的。该模型首先将输入的图像经过卷积层、池化层中的相关、重复操作，将图像中的关键信息提取成图像特征，并用这些提取出来的关键的特征对图像进行分类或者识别图像中的目标。从上文的论述中可以看出，卷积神经网络在图像的分类与图像目标识别的任务中训练模型阶段与与测试阶段不需要人工参与，并且具有计算速度快（参数共享、降采样等级制）鲁棒性强的优点。目前，高精度的图像目标识别以及图片分类任务，卷积神经网络已经成为最重要的关键技术[16]。传统的图像识别技术需要利用图像特征描述子SIFT提取图像的特征，然而该技术具有提取特征不准确并且时间开销比较大的缺点。基于上述问题，研究者相继提出了一些列的改进算法：[17-19]等。传统的图像分类与图像目标识别的技术利用描述子提取了大量的特征之后，需要用其它技术手段对提取的特征进行分类，即简历图像的视觉词典，然后在视觉词典的基础上，计算图像的特征与词典中视觉单词的距离，得到与目标距离最近的视觉单词，最后统计图像中视觉单词的数目，并以直方图的形式表现出来。但是该方式不能提取图像的抽象含义并且在模型搭建的过程中需要经验丰富的工程师利用大量的领域知识进行建模，不仅速度慢，准确率也不高。研究者为了得到更加精准的图像特征，提出了一系列的技术，包括再统计、编码等[20-22]。1997年Joachims等人认为文档中词语与词语之间不具有关联性，将文档当成一个纯粹的填装词语的袋子，提出了词袋模型(BoW)。2003年Sivic等人将图像当做特殊的“文档”,用来进行图像分类与图像目标的识别。该模型首先提取感兴趣的关键词，然后将文档表示成与顺序无关的关键词组合，最后利用统计的方法对文档中关键词出现的频率进行计算。该模型具有简单、易于理解的特点，研究者们将该模型成功地应用到自然图像的目标识别中并取得了较好的结果[16,17]，首先将图像中多个局部特征转化为视觉单词，然后将输入的每一幅图像表示为视觉单词的直方图。将词袋模型应用到图像的目标识别中主要包括提取和描述特征，创建字典，量化特征等过程。随后，大量的研究者在词袋模型的基础上做了改进：Lazebni等人在词袋模型的基础上引入了金字塔模型[23]；Yang等人在Lazebni的工作基础上增加了稀疏编码理论，提出了基于稀疏编码的金字塔模型。

* + 1. 卷积神经网络在图像识别中的发展

2014年Girshick等人在卷积神经网络的基础上，增加了线性回归、支持向量机等技术，提出了一种全新的卷积神经网络架构R-CNN，该结构是第一个将卷积神经网络运用到图像的目标识别中。传统的图像识别技术需要利用图像特征描述子SIFT提取图像的特征，然而该技术具有提取特征不准确并且时间开销比较大的缺点。在模型训练阶段人工的模型参数调优的方式，极大地限制了传统图像分类以及图像目标识别技术的准确性。R-CNN模型不需要人工提取图像特征，并且参数的调整过程是完全自动化的。该模型首先将输入的图像经过卷积层、池化层中的相关、重复操作，将图像中的关键信息提取成图像特征，并用这些提取出来的关键的特征对图像进行分类或者识别图像中的目标，具有高度的自动化、识别准确度高的特点，很快成为目前图像目标识别领域的主流技术。

基于卷积神经网络的图像目标识别技术首先将输入的图像划分分区，即将输入的图像划分为不同的兴趣区，然后将兴趣区进行裁剪并输入到卷积神经网络中，接着在卷积层与池化层提取高纬度、抽象的图像特征，最后利用支持向量机对兴趣区进行分类。在该类模型中兴趣区的分类结果视为目标的类别，对应的位置Wie目标位置，相关的算法包括：R-CNN、Fast R-CNN和Taster R-CNN等。R-CNN在对输入的图像进行兴趣区划分时，通常产生2000多个区域，极大地增加了模型的训练与测试开销。在R-CNN的基础上，Fast R-CNN在R-CNN的基础上增加了区域池化层，减少了大量的重复计算，大大降低了模型训练与测试的时间。

* 1. 小结

本章首先介绍了神经网络的基本概念、开源框架，然后介绍

了卷积神经网络、目标识别以及卷积神经网络在图像识别领域中的研究工作。

3 基于文本挖掘的神经网络企业信用评价方法

3.1 动机

企业信用评级是指信用评级机构利用某种方法或者模型对制造业企业、服务业企业、建筑业企业的信用进行评级。随着评级方法的不断完善，企业信用评级方法考察的范围越来越广：服务能力、发展潜力、技术能力、战略眼光等。

目前投资者与企业之间存在严重的信息不对称的情况，即企业内部对自己有深刻的认知，然而投资者缺少企业的相关信息，或者需要从海量的数据中寻找企业的相关信息，这种情况不仅为投资者投资企业增加了难度，并且增加了投资者自身的收益风险。

当下社会处于经济转型期，很多经营者与投资者依然存在“旧思想”和“旧观念”，即国家政府掌控企业的资金。另一方面，我国的企业评级起步较晚，没有受到广泛的普及。然而，企业信用普及是一种最直观有效的信息，为投资者投资企业提供了重要的参考。传统的企业信用评级方法利用的信息之间往往存在复杂的非线性关系，因此传统的企业评级技术不能很好的适应当前的企业信用系统。

近年来，随着企业评级技术与计算机硬件与软件的不断发展，企业各方面的数据呈指数增长，使得企业评级的准确度与参考价值得到了明显地提高。然而，数据量的增加，也为企业信用评级带来了一些列的问题，具体来说：（1）企业的信用度与评价指标之间呈现出非线性关系。传统的企业评级的线性模型不再适用于当前的企业信用评级系统，需要一种能够反映属性之间复杂关系的模型；（2）参考指标多样化。传统的企业评级的综合模型在此情况下，即便是领域专家也不可能了解每一个属性，并且人工的方式易造成不公平的现象。因此，提出可靠高效的企业信用评级方法是一个亟待解决的重要问题。

神经网络模型在当前时代背景下得到了广泛的发展，例如，运用神经网络在高分辨率遥感图像的识别任务上，世界上最先进的高分辨率遥感图像的分辨率达到了分米级，合成孔径雷达（Synthetic Aperture Radar, SAR）已达米级，采用干涉测量技术的合成孔径雷达（Interferometric Synthetic Aperture Radar, InSAR）相对位移精度达到了毫米级，高光谱图像（Hyperspectral Image）的分辨率甚至达到了纳米级。另一方面，卫星重访周期的单位已缩短至天。上述特征使得拍摄的高分辨率遥感图像不仅具备数量多、多源化、异构化的特征，还具备高维度、多尺度、非平稳的内部特性，具备大数据的特点[1]。然而，适配于中低分辨率的传统遥感影像解译方法，在识别包含车辆、飞机、轮船等复杂特征的高分辨率遥感影像时效率低下，甚至无法准确识别。面对新时代高分辨率遥感影像目标识别的更高要求，亟需探索全新的、更加高效的高分辨率遥感影像目标识别方法与模型，用来提高高分辨率遥感图像的目标识别精度、识别技术的智能化水平[2-4]。很多工作表明神经网络在很多领域表现突出甚至达到了人类的水平。

传统的企业信用评级方法，不能适用于现在企业的评级体系，具体来说：（1）判别分析法：该方法首先从已有的数据中区分正常企业与非正常企业，然后再从分类的数据中提取特征，并总结出一个函数作为系统的判别函数。该方法不适合应用在企业属性之中出现非线性关系的情形；（2）综合评定方法：该方法首先分析企业设计到的属性，然后为每一个属性进行打分，接着利用某种算法（加权平均法）计算出企业的总得分。由于该方法涉及到专业人员评分，因此存在不平等的现象、

在智能化时代的背景下，云计算为企业信息提供了存储空间，使得海量的企业信息存储成为可能。研究者可以从大量存储的数据中识别、提取感兴趣的与企业信用相关的信息，通过分析判别方法、简历模型等方法建立信用评价系统。

基于上述分析，从海量的数据中提取企业信用相关的信息和信用评价方法决定了企业信用评级的可靠性。在当前的智能时代下，企业信息具有数量多、种类多、隐藏性高等特征，具体来说：（1）在数量方面：经济的全球化，导致企业与外界的交流更加密切，这也导致了交易数据增长呈现出几何趋势。公司为了自身发展注重数据的收集，依据收集的数据进行对比和分析为下一次的决策提供有效的帮助，并且投资者为了降低风险对公司进项全面的了解。上述现象导致企业数据量十分巨大；（2）在种类方面：随着公司的不断发展，公司的业务呈现出多元化的特点，相应的数据也呈现出种类多的特点；（3）在隐藏性方面：大量的数据在为研究者提供大量感兴趣的内容的同时也为提取有价值的数据或者感兴趣的数据增加了难度。

机器学习与神经网络在文本处理以及目标识别领域具有优秀的表现[5]。另一方面，随机计算机技术以及硬件的不断发展，实现了对大量数据进行操作，这使得神经网络运用到现代企业信用评级成为可能。当前企业信息呈现出的特点增加了提取有关数据的难度，该数据作为神经网络的输入训练模型、预测企业的信用分类。如何正确地将神经网络运用到企业信用等级划分中，自动、高效地分类目标具有重要的研究价值。

3.2 基于文本挖掘的神经网络企业信用评级框架（MBP-CCS）

本文基于神经网络的基本流程与企业信用评级的基本原理，提出了基于文本挖掘的神经网络的企业信用评级框架（图3-1）。在将企业数据输入到神经网络之间，需要对企业信用数据做以下处理：数据的预处理、分词和关键词提取。然后，基于神经网络的基本原理，在对企业信用评级的过程中，首先通过文本挖掘得到的数据进行预处理，该步骤主要是将数据处理成神经网络可以识别的范围；然后对规则的数据进行中心化，该步骤的主要作用是使得数据呈现正态化；最后输入到模型中，并且利用误差反向传播的算法更新权重。

在智能化时代企业信息呈现出数量多、种类多、隐藏性高等特征，具体来说：（1）在数量方面：经济的全球化，导致企业与外界的交流更加密切，这也导致了交易数据增长呈现出几何趋势。公司为了自身发展注重数据的收集，依据收集的数据进行对比和分析为下一次的决策提供有效的帮助，并且投资者为了降低风险对公司进项全面的了解。上述现象导致企业数据量十分巨大；（2）在种类方面：随着公司的不断发展，公司的业务呈现出多元化的特点，相应的数据也呈现出种类多的特点；（3）在隐藏性方面：大量的数据在为研究者提供大量感兴趣的内容的同时也为提取有价值的数据或者感兴趣的数据增加了难度。接下来对架构中的每一部分做详细解释：

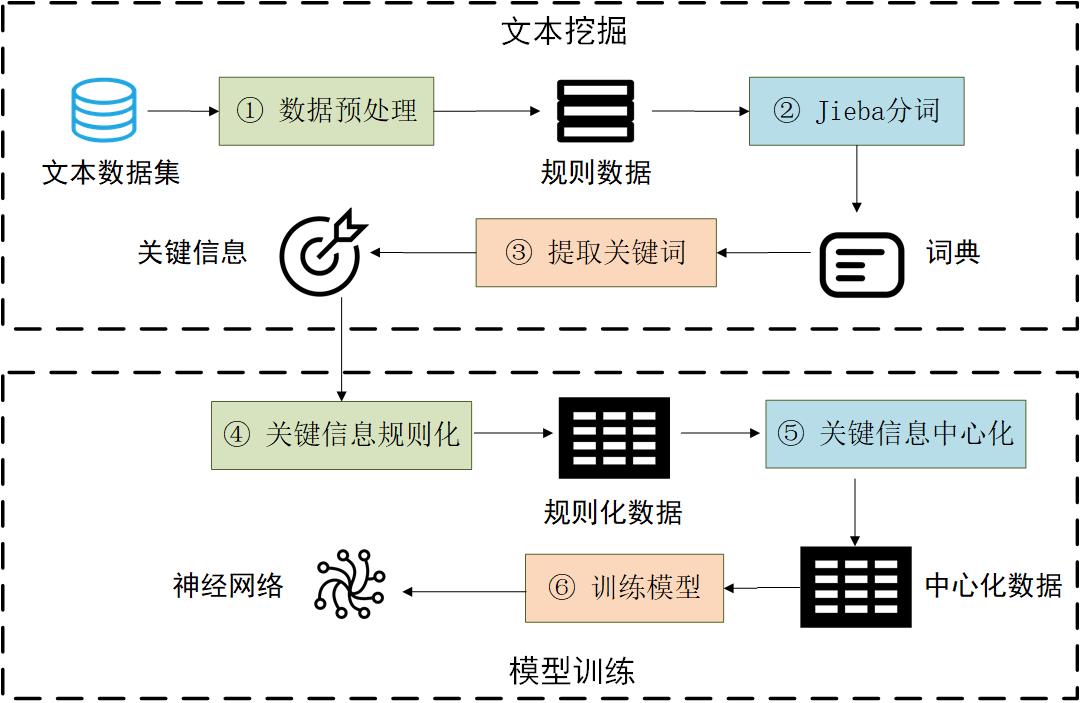


图3-1 基于文本挖掘的神经网络企业信用评级模型

Figure 3-1 Neural Network Enterprise Credit Rating Model Based on Text Mining

**数据预处理：**从网上获取数据的方式有多种：（1）设计爬虫爬取数据；（2）基于某些公开的接口获取数据；（3）从数据库中下载数据等。由于获取的数据呈现出格式混乱、形式不统一、大量信息重复、含有噪声和纬度高的特点，并不能直接用于知识发现，需要对获取的数据进行预处理。数据预处理的方法一般有：数据清理、数据集成、数据规约和数据变换，具体来说：（1）在数据清理方面：该步骤主要是补全获取的数据的空白项，光滑噪声数据，利用聚类技术识别离群点，纠正数据中的不一致现象（例如：在属性名称为性别的数据中，存在字符串数字等来表示性别）。其中补全空白项的方式一般有：忽略元组法，即将存在空白项的元组删掉；人工填补法：研究者利用以往的经验填写数据，该方法需要研究者对数据有较深的认识；随机数发，即根据当前空缺元组的数据范围随机一个数值填补。噪声是待测对象的随机误差或综合方差，研究者可以利用线性回归或者非线性回归对获取的数据进行拟合，护着利用聚类的方法识别利群点。数据集成主要应用在从多个数据源中回去数据的情况。在这种情况下，数据的不一致情况严重，数据冗余度较高。该步骤直接影响后期数据挖掘与分析的准确度。在存在多元数据的情况下，研究者首先需要处理实体识别问题，即将多个数据源的属性进行关联。建立属性的关联之后，要对属性的冗余性与相关性及进行分析。如果存在一个属性，该属性可以从另一个属性中导出，则该属性被视为冗余属性。数据规约主要是为了保持原数据信息量的同时，进一步减小数据量节约后期训练模型、测试模型所需要的时间。数据规约的方法有多种，具体来说：（1）数据压缩：该方法一般通过变换的形式，得到原有的数据压缩数据，在使用的过程中，在进行还原，如果还原后的数据与初始数据一致则称该压缩为无损压缩，否则成为有损压缩；（2）数量规约：该方法主要利用其它形式的数据（例如较小的数据）来代替原有的数据；（3）维规约：该方法主要是利用已有的技术（主成分分析）降低属性中数据的维度，去除冗余变量和不相关变量，降低计算时间，提高模型的分类准确度。

**Jieba分词：**该框架是是被广泛使用的分词框架，支持三种分词模式：（1）精确模式，在该模式下输入的文本将被最细致的分开。该模式经常用于文本分析；（2）全模式：该模式将词作为目标，扫描文本中的词语；（3）搜索引擎模式，

该模式建立在全模式的基础上，对全模式切分的词进行筛分，识别字数较多的词语，再进行切分。本文利用该框架对获得数据进行分词。

**提取关键词：**本文依据[24]中提到的企业信用评级指标（图3-2），提炼本文的关键词，并获取每个关键词相应的信息。图3-2中，偿债能力是指企业偿还债

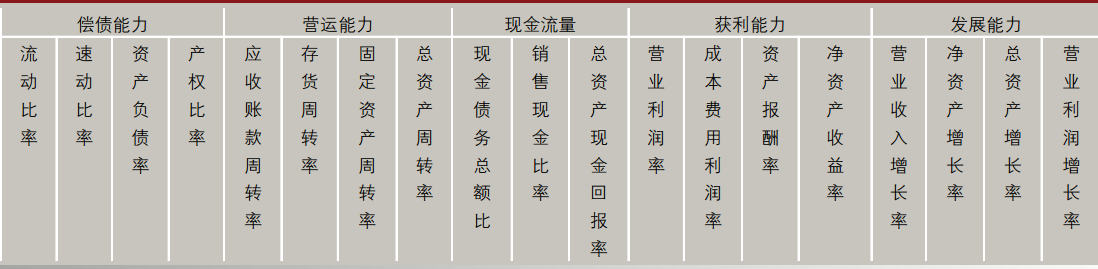
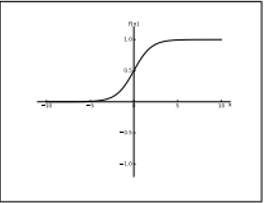


图3-2 企业信用评价指标

Figure 3-2 Corporate credit evaluation index

务的能力，可以细分为：流动比率、速动比率、资产负债率以及产权比率；运营能力是指企业根据自己的产业在市场中的效率与效益，可以细分为：应收账款周转率、存货周转率、固定资产周转率和总资产周转率；现金流量是指企业在投资某个项目的整个周期中现金的流入流出的总额，可以细分为：现金债务总额比、销售现金比率和总资产现金回报率；获利能力是指企业利用产出转化为金钱的能力，可以细分为：营业利润率、成本费用利润率、资产报酬率和净资产收益率；发展能力是指企业扩发规模、壮大实力的潜在能力，对企业的进一步发展至关重要，可以细分为营业收入的增长率，净资产收入增长率、总资产增长率和营业利润增长率。本文依据上述指标从企业的大量数据中挖掘相关信息，然后再对信息进一步处理，输入到决策模型中。详细地处理步骤在接下来的章节种描述。

**关键信息规则化：**在2.1.1章节中，本文讨论了目前神经网络中常用的阈值函数。本文主要是利用自适应神经网络模型（SOM）对企业进行自动分类。SOM的阈值函数为“S”函数，该函数的取值范围是，其函数图像展示在图3-3中。

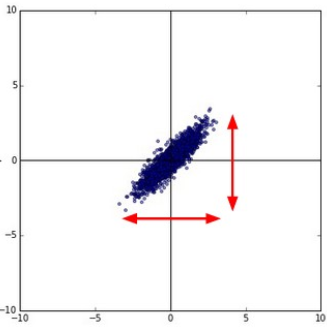
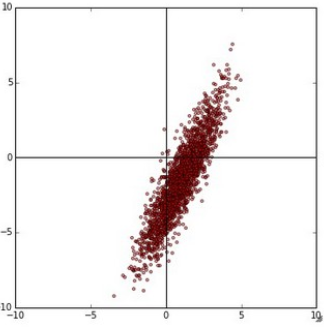
****

**图3-3 “S”函数图像**

**Figure 3-3 the Figure of S**

从图中可以看出，当输入额值特别大或者特别小时，函数值的差别不大，因此为了区别不同输入值对应输入值的不同，一般在自适应神经网络中要求输入值的取值范围在或者。另一方面，神经网络中的输入一般是数字，因此在存在字符串的情况下，需要对字符串与数字用之间建立映射关系，将字符转化为数字。

**关键信息中心化：**信息中心化也成为归一化，即在规则化的基础上对数据进行操作，使得每一个属性的值处于之间。图3-4展示了普通数据与归一化



（a）普通数据 （b）归一化数据

图3-4 数据归一化前后对比

Figure 3-4 Data before and after normalization

后的数据，从图中可以明显看出，归一化后的数据更集中，聚合度更高。

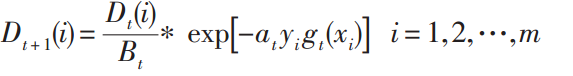
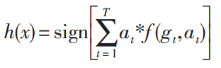
**训练模型：**本文利用神经网络中无监督的自适应性神经网络（SOM）作为模型，通过训练该模型来作为企业信用度的弱分类器，然后利用adaboost算法对该分类器进行升级，得到分类能力更强的分类器，具体来说：

1. 训练模型：从样本空间中随机选取一定数目的数据来训练模型，剩下的数

据作为测试数据。在训练的过程中，SOM神经网络的参数通过随机的方式进行初始化，在学习的过程总不断更新权重。

1. 输出测试结果：将测试数据输入到训练好的模型中，收集输出的结果与预期的结果并将输出的结果与预期结果与实际结果的误差作为序列。
2. 计算权重：基于到的序列，利用一下公司计算预测序列的权重。

其中，是第次输出结果的误差，是第次结果的权重。

1. 依据公式得到下一轮样本的权重。
2. 基于以下公式得到强分类函数：。

**4 经验研究**

为了评估基于本文挖掘的神经网络企业信用评级模型的效率、有效性以及准确率，本章构造了经验研究来评价提出的模型。

* 1. 研究问题

本章实验围绕以下三个问题展开讨论：

1. 基于本文挖掘的神经网络企业信用评级框架划分企业信用的准确率。本文利用上海证券交易所A股市场上上市的制造业为研究对象，选取了50家正常的企业，即信用度良好，10家不正常的企业，即信用度低的企业来训练模型，最后分别选取10家正常和不正常的企业进行测试，验证模型的准确性。
2. 基于本文挖掘的神经网络企业信用评级框架划分企业信用的性能。本文利用识别一家企业信用的时间为标准衡量本文提出的方法的时间开销。
   1. 实验对象

本文利用中国上市公司（上海证券交易所）网站信息，首先在可获得的企业列表中筛选制造业企业，然后随机选取200家“正常营业”的企业和50家“非正常”企业。为了便于研究，本文将上市公司被实施特别处理的年份记为，根据中国上市公司的年报披露制度采用 和 年的财务数据作为样本数据，最终形成了容量为的面板数据。

为了保证统计的可靠性，本文胸可获取的企业列表中随机10次，每次取200家正常营业的企业，50家非正常营业的企业。去每次分类准确度的均值。

* 1. 实验设置

本节主要介绍本文实验相关的设计细节。

* + 1. 待评估的技术

本文提出的MBP-CCS作为研究的对象。为了突出基于文本挖掘的adaboost神经网络对企业信用度评估的有效性与准确率，本文评估与比较MBP-CCS和传统神经网络（BP）的有效性与效率。

* + 1. 度量标准

在机器学习领域中，分类和预测是主要的研究问题。然而任务场景的多样性导致以某一个技术适用于所有的情形。为了判断研究者所提技术的有效性，以下几个度量标准受到广泛的应用：（1）准确率：技术识别的目标中正确的目标占的比例；（2）召回率：识别的目标种类占所有目标种类的比率；（3）平均准确率：反应了某一个目标的全局性能；（4）平均准确率均值：反应了多个目标的全局性能。

对于研究问题（1），本文利用各个类别的平均准确率和所有类别的平均准确率均来衡量。

对于研究问题（2），本文利用平均准确率来衡量所提框架的有效性。此外，目标识别所需时间是衡量目标识别技术的重要指标，本文提出一种度量指标：T-measure，表示从一组数据中识别出目标的分类所需的时间。

* + 1. 实验环境

本文使用 python 语言编写测试脚本， 运行在 Windos10 64 位操作系统上。该系统有 4 个 CPU 和 16GB 内存。

* + 1. 实验流程

本文的实验主要分为两个：基于MBP-CCS和BP的企业信用度分类的准确率。接下来详细描述具体的实验流程。

**数据预处理：**本文从网上下载了大量的企业信息。获取的数据呈现出格式混乱、形式不统一、大量信息重复、含有噪声和纬度高的特点，并不能直接用于知识发现，需要对获取的数据进行预处理。数据预处理的方法一般有：数据清理、数据集成、数据规约和数据变换，具体来说：（1）在数据清理方面：该步骤主要是补全获取的数据的空白项，光滑噪声数据，利用聚类技术识别离群点，纠正数据中的不一致现象（例如：在属性名称为性别的数据中，存在字符串数字等来表示性别）。其中补全空白项的方式一般有：忽略元组法，即将存在空白项的元组删掉；人工填补法：研究者利用以往的经验填写数据，该方法需要研究者对数据有较深的认识；随机数发，即根据当前空缺元组的数据范围随机一个数值填补。噪声是待测对象的随机误差或综合方差，研究者可以利用线性回归或者非线性回归对获取的数据进行拟合，护着利用聚类的方法识别利群点。数据集成主要应用在从多个数据源中回去数据的情况。在这种情况下，数据的不一致情况严重，数据冗余度较高。该步骤直接影响后期数据挖掘与分析的准确度。在存在多元数据的情况下，研究者首先需要处理实体识别问题，即将多个数据源的属性进行关联。建立属性的关联之后，要对属性的冗余性与相关性及进行分析。如果存在一个属性，该属性可以从另一个属性中导出，则该属性被视为冗余属性。数据规约主要是为了保持原数据信息量的同时，进一步减小数据量节约后期训练模型、测试模型所需要的时间。数据规约的方法有多种，具体来说：（1）数据压缩：该方法一般通过变换的形式，得到原有的数据压缩数据，在使用的过程中，在进行还原，如果还原后的数据与初始数据一致则称该压缩为无损压缩，否则成为有损压缩；（2）数量规约：该方法主要利用其它形式的数据（例如较小的数据）来代替原有的数据；（3）维规约：该方法主要是利用已有的技术（主成分分析）降低属性中数据的维度，去除冗余变量和不相关变量，降低计算时间，提高模型的分类准确度。

**Jieba分词：**该框架是是被广泛使用的分词框架，支持三种分词模式：（1）精确模式，在该模式下输入的文本将被最细致的分开。该模式经常用于文本分析；（2）全模式：该模式将词作为目标，扫描文本中的词语；（3）搜索引擎模式，

该模式建立在全模式的基础上，对全模式切分的词进行筛分，识别字数较多的词语，再进行切分。本文利用该框架对获得数据进行分词。

**提取关键词：**本文依据[24]中提到的企业信用评级指标（图3-2），提炼本文的关键词，并获取每个关键词相应的信息。

**关键信息规则化：**在上一步的基础上，将字符型数据转化为浮点型数据。

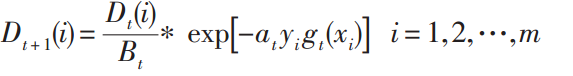
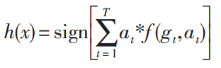
**训练模型：**本文利用神经网络中无监督的自适应性神经网络（SOM）作为模型，通过训练该模型来作为企业信用度的弱分类器，然后利用adaboost算法对该分类器进行升级，得到分类能力更强的分类器，具体来说：

1. 训练模型：从样本空间中随机选取一定数目的数据来训练模型，剩下的数

据作为测试数据。在训练的过程中，SOM神经网络的参数通过随机的方式进行初始化，在学习的过程总不断更新权重。

1. 输出测试结果：将测试数据输入到训练好的模型中，收集输出的结果与预期的结果并将输出的结果与预期结果与实际结果的误差作为序列。
2. 计算权重：基于到的序列，利用一下公司计算预测序列的权重。

其中，是第次输出结果的误差，是第次结果的权重。

1. 依据公式得到下一轮样本的权重。
2. 基于以下公式得到强分类函数：。
   * 1. 潜在的风险分析

影响本文实验结果的一个潜在风险是测试脚本实现的正确性。 明显地，不正确的测试脚本直接影响实验结果的正确性。测试脚本经过不同开发人员检查、修正， 因此相关测试技术的实验脚本都是正确的。另一个影响实验结果的风险是度量指标（平均准确率和T-measure） 的可靠性。 单次实验结果不能有效地反映度量指标的真实值，本文重复 5 次实验保证度量指标在统计上的可靠性。

1. 实验结果与分析

本节展示经验研究的实验结果。

* 1. 数据预处理结果
     1. Mann- Whitney U 检验

为了检验从网上爬取的数据的正确性，以及正常营业的企业与非正常营业的

企业从统计学的角度存在真实的差异，本文利用Mann-Whitney U检验技术，对获取的数据进行秩和检查，部分结果展示在表5-1中。

表5-1 Mann- Whitney U 检验结果

Table5-1 the results of Mann- Whitney U

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 收款转账率 | 存货周转率 | 收入增长率 | 利润增长率 | 销售现金比 |
| Mann-Whitney U | 1171.000 | 1423.000 | 1321.000 | 1407.000 | 1178.000 |
| Wilcoxon W | 1636.000 | 6473.000 | 1786.000 | 1872.000 | 1643.000 |
| Asymp.Sig. | 0.069 | 0. 670 | 0.323 | 0.607 | 0.0755 |

从表中可以看出，表5-1中的对象在显著水平为0.05的情况下，均不能显著的区分不同公司的信用水平。因此，本文在后续的实验过程中将上述属性删掉。

* + 1. 主成分分析

原始数据的维数较高，并且包含大量的属性。无关的属性值将影响最终的分类结果。因此，本文利用主成分分析的方法，选择与分类结果紧密相关的属性，删除无关属性，不仅简化了模型，提高了分类的准确率，而且减少了模型的训练时间与验证时间。本文利用SPASS得到数据集中不同属性的贡献率，结果展示在图5-1中。

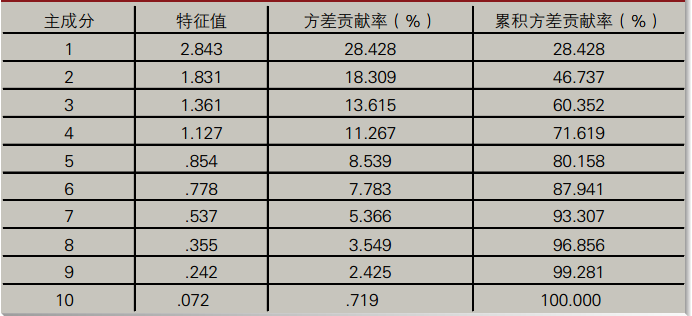


图5-1 不同属性的贡献率

根据图5-1，本文利用贡献率达到80%的原则，得到最重要的5个属性：获利能力、偿债能力、营运能力、成长能力和自有资产管理能力。基于这5个属性的数据训练网络。

* 1. 模型训练结果

通过数据训练的SOM可视化模型图如图5-2。图中展示了正常企业与非正常

企业对应的激活函数。其中“o”表示正常企业，“◇”表示非正常企业。

* 1. 测试结果

模型分类的准确度是衡量模型实用性的重要指标，本文将测试数据分别输入到训练好的两个模型中，得到的准确率均值记录在表5-2，每一次实验结果的分布图如图5-3。从表5-2中可以看出，本文所提的MBP-CCS方法相对于传统的方法有较大的改进，并且从图5-3可以看出，本文提出的方法在每一次的随机测试过程中表现稳定，具有较高的分类结果。

表5-2 测试结果

Table 5-2 the results of testing

|  |  |
| --- | --- |
| 方法 | 准确率 |
| 传统的BP神经网络 | 72.93% |
| 基于Adaboost的BP神经网络 | 81.34% |

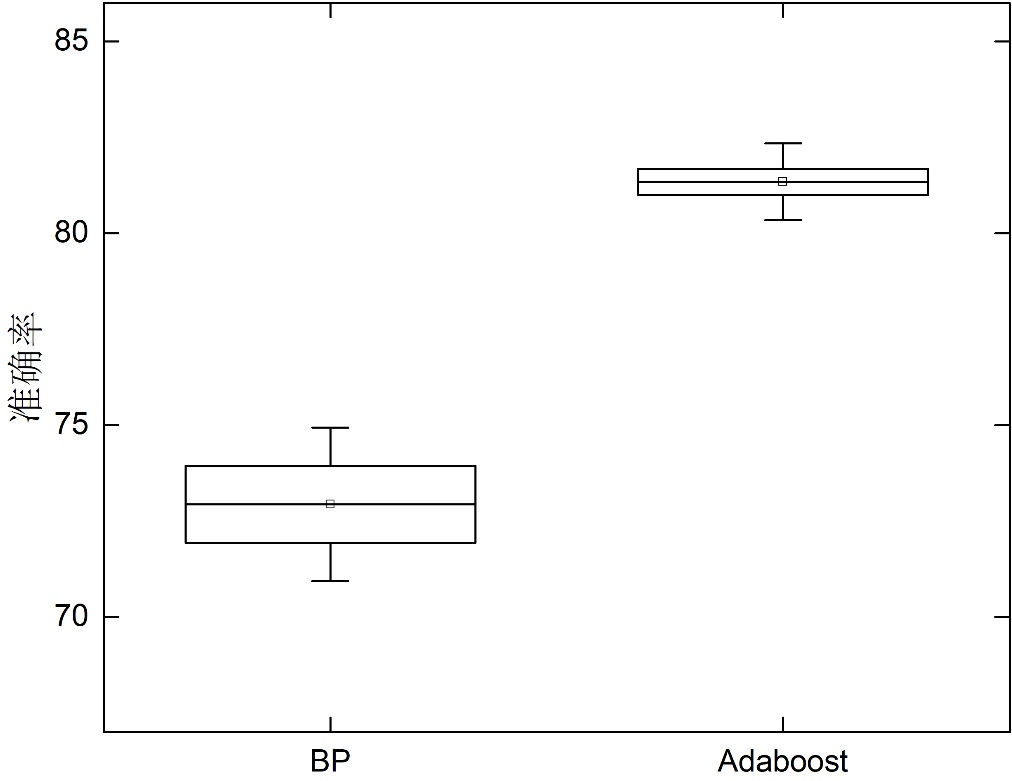


图5-3 模型测试结果分布图

5.4 时间开销

本文在传统的BP神经网络的基础上增加了Adaboost算法，将原来的弱分类器变为强分类器，加强了BP的分类能力，从实验结果看，本文提出的MBP-CCS的确具有较高的分类结果。但是该方法引入了额外的计算，为此本文统计每一次随机测试过程中的时间开销，并计算平均值展示在表5-3中。从表中可以看出，本文所提的方法的确引入了额外的时间开销，但是较之传统的方法多用了3.08s，处于可以接受的范围内。

表5-3 测试的时间开销

Table5-3 The overhand of testing

|  |  |
| --- | --- |
| 方法 | 时间（s） |
| 传统的BP神经网络 | 12.24 |
| 基于Adaboost的BP神经网络 | 15.32 |

1. 工作总结与展望

企业信用评级是目前研究的热点领域，具有重要的理论研究意义与广泛的应用价值。面对当前投资人与企业信息的不对称现象，企业信用评级成为投资者了重要的参考标准。然而在当前信息爆炸的时代，从大量数据中挖掘、提取有价值的信息是一件具有挑战的事情。并且传统的企业评估技术无法应对多变的评价指标。

神经网络是机器学习的一个重要分支，在实际应用中表现良好甚至达到了人类的水平。神经网络作为一种广为研究与应用的学习模型，被认为是最强大的分类与目标识别模型。

本文旨在利用自适应神经网络（SOM）自组织的无监督学习能力自动地对企业进行分类，在本文挖掘的基础上提出一种adaboost与SOM结合的企业信用评级框架，且通过经验研究的方式验证所提框架的正确性与有效性。具体来说，本文的主要研究内容包括：

1. **提出了基于文本挖掘的adahoost和神经网络企业评级框架(****MBP-CCS)**：**MBP-CCN**利用文本挖掘相关的技术，得到相应的数据，训练自适应神经网络（SOM），然后利用adahoost算法得到强分类器。
2. **采用经验研究验证所提模型对企业信用评级分类的准确性以及有效性**：利用中国上市公司网站的信息，随机选取200个正常营业的制造业公司和50个非正常营业的制造业公司来训练模型，并利用测试用例集评估所提模型的正确性与准确性。

本文的所提的基于文本挖掘的企业评级**MBP-CCN**模型，改进了传统神经网络的企业信用评级方法，并且**MBP-CCN**在测试的过程中表现突出，具有很大的研究潜力。