(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108694476 A
(43)申请公布日 2018.10.23

(21)申请号 201810700770.5

(22)申请日 2018.06.29

(71)申请人 山东财经大学
地址 250014 山东省济南市历下区二环东路7366号山东财经大学燕山校区

(72)发明人 王玉洁 刘慧 张彩明 郭强 刘鑫

(74)专利代理机构 济南舜昊专利代理事务所（特殊普通合伙）37249
代理人 侯绪军

(51)Int.Cl.
G06Q 10/04(2012.01)
G06Q 30/02(2012.01)
G06Q 40/04(2012.01)

权利要求书2页 说明书11页 附图2页

(54)发明名称

一种结合财经新闻的卷积神经网络股票价格波动预测方法

(57)摘要

本发明提供一种结合财经新闻的卷积神经网络股票价格波动预测方法，本发明利用自然语言处理技术来提取相关新闻中的特征，从而分析、观测财经新闻与股票价格走势的关联程度。本文结合新闻报道的有效信息，提出了一种基于卷积神经网络的股票价格波动预测方法。首先，将新闻分词，并提取主要事件，利用出现次数最多的前3000个词语作为关键词，并使用Glove模型将其表示为低维稠密的词向量；其次，将新闻特征和股票价格对应起来，把时间划分成短、中、长三个时间段，用卷积神经网络来模拟新闻事件对股票价格变动的短期和长期影响；最后，通过训练好的模型预测股票的涨跌情况。

CN 108694476 A

1.一种结合财经新闻的卷积神经网络股票价格波动预测方法，其特征在于，方法包括：

步骤一：扫描语料库，预设关键词，设定扫描窗口长度，在设定的扫描窗口内配置共现矩阵；

步骤二：基于共现矩阵配置共现概率；

步骤三：将语料库中词语转化为索引矩阵，设定要保留的关键词个数，设定模型参数，基于共现概率定义词向量矩阵；将数据集中的数据切分成训练集和测试集两部分，利用索引与训练出来的关键词矩阵一一对应，获得数据集的词向量；

步骤四：基于预测股价的涨跌作为一个分类问题，即涨和跌两类；将训练集和测试集的词向量矩阵载入卷积神经网络，新闻词向量和股票利用日期信息一一对应，利用股价的涨跌作为特征标签，将训练集和测试集分别提取出特征标签和类别标签；

步骤五：预测模型的网络结构为5层，其中前两层为卷积层，第一层神经元个数为64个，卷积核大小为3*100，输入矩阵大小30*100，激活函数为ReLU函数，第二层卷积神经元个数为32个，卷积核大小为3*50，第三层为最大池化层，池化大小为18*1，利用18*1的窗口扫过矩阵，提取每个窗口中的最大值；得到一个新矩阵；

第四层和第五层为全连接层，第四层的全连接层神经单元节点数为64，激活函数为ReLU函数；

第五层激活函数则是softmax，是进行最后的分类，得到了每只股票后一天股价的涨跌情况。

2.根据权利要求1所述的结合财经新闻的卷积神经网络股票价格波动预测方法，其特征在于，

步骤一还包括：

设定窗口长度为n；对设定窗口中的句子进行扫描，得到关键词i，j在设定窗口中出现的次数X_{ij}.遍历整个语料库后得到共现矩阵X。

3.根据权利要求1所述的结合财经新闻的卷积神经网络股票价格波动预测方法，其特征在于，

步骤二还包括：计算关键词i，j的共现概率

$$P_{ij}=P(j|i)=\frac{X_{ij}}{X_i}, \qquad (1)$$

这个比率反映了词语之间的相关性，称词i和j分别为中心词和背景词，接下来是利用模型训练词向量，使用词向量表达共现概率的比值，任意一个比值需要三个单词i，j，k：

$$F(w_i,w_j,w_k)=\frac{P_{ik}}{P_{jk}}, \qquad (2)$$

这里，w是单词的词向量，P_{ik}是单词k出现在单词i上下文的概率，P_{jk}是单词k出现在单词j上下文的概率.$F(w_i,w_j,w_k)$是关于词向量的函数，由于向量的本质是线性结构，式(2)中F的形式是F＝exp，因此可以推出词向量需要满足的等式：

$$\log(X_{i,j})=v_i^T v_j+b_i+b_j, \qquad (3)$$

损失函数为：

$$J=\sum_{i,j=1}^{V} f(x_{ij})(\omega_i^T\omega_j+b_i+b_j-\log x_{i,j})^2, \quad (4)$$

V是词汇量的大小，f(x)是权重函数，b_i和b_j是偏置项，权重函数f(x)可以参数化为：

$$f(x)=\begin{cases}(x/x_{max})^{\alpha} & \text{if } x<x_{max}\\ 1 & \text{otherwise}\end{cases}, \quad (5)$$

根据经验值，α等于$\frac{3}{4}$；

得到全部词向量之后，每个词向量都是中心词和背景词向量之和。

4.根据权利要求1所述的结合财经新闻的卷积神经网络股票价格波动预测方法，其特征在于，

步骤三中：预设保留的关键词个数为3000至4000个。

一种结合财经新闻的卷积神经网络股票价格波动预测方法

技术领域

[0001] 本发明涉及大数据领域，尤其涉及一种结合财经新闻的卷积神经网络股票价格波动预测方法。

背景技术

[0002] 股票市场因其相对投资灵活的操作特征，已经成为了金融市场的重要组成部分。股票市场高风险与高回报的特性吸引了很多经济学家和投资爱好者，但在一般情况下，股民很少能够准确判断股票市场的价格变动情况。因此，研究并构建一个科学、预测精度高的模型，以有效把握股票市场波动规律，帮助投资者规避风险、提高收益，具有重要的理论意义和现实意义。

[0003] 对于股票的收益，我们的期望是回报率最大的同时风险最小，这就需要掌握一种有效的方法来分析股票价格交易的波动规律和走势。到目前为止，针对股票价格波动规律，国内外研究学者提出了很多不同的理论方法，分为以下几种：

[0004] (一)时间序列分析方法

[0005] 时间序列分析是一种典型的根据时间递进关系进行预测的方法，其主要思想是利用已有的与时间序列相关的数学模型和经济行为作为研究对象，在时间序列中发现某段时间内经济商品价格的波动规律，从而进一步预测商品的发展规律。而股票也可以作为一种商品，在没有重大事件影响的情况下，股价每天都表现出一定的规律，这和时间序列的思想吻合，它体现了股票价格运行的长期趋势。典型的时间序列分析方法主要有指数平滑模型、自回归条件异方差模型等。

[0006] (二)灰色预测法

[0007] 灰色理论认为，尽管系统的行为现象无法清楚确定，但数据是有序的，且有整体功能。而灰数，就是在杂乱中寻找规律。灰色预测法是一种对含有不确定因素系统进行预测的方法，通过对系统因素的关联分析，生成有较强规律性的数据序列，建立对应的微分方程，从而预测事物未来的发展趋势。在股票系统中引起股票涨跌的不确定因素有很多，因此可以利用灰色预测方法预测股票价格的波动情况。徐维维等人利用灰色系统理论建立了GM(1,1)模型，对所得结果进行残差修正以计算出股票价格。实验表明，利用灰色理论具有较高的精度，但是只适用于短期的股指数据。

[0008] (三)证券投资分析方法

[0009] 证券投资分析方法主要分为两种：一种是技术分析方法，另一种是基本分析法。技术分析法主要是根据股票价格波动规律的历史数据资料进行分析和研究，在此基础上预测股票价格未来的波动趋势。基本分析法则关注每只股票交易价格中固有价值的波动情况，通过价值波动理论对上市公司的经营情况、资金储备和资金链运转情况进行详细研究，从而根据股票内在价值的波动情况获得股票价格的波动规律。

[0010] (四)人工神经网络

[0011] 人工神经网络是一种非线性模型，它充分利用训练数据，经过多次迭代学习原始

特征空间中训练数据之间的内在联系，构建一个具有良好的非线性逼近能力的学习模型。常用的神经网络主要有反向传播算法(BP)神经网络、回归神经网络(RNN)、深度信念网络(DBN)等。因此，对于像股票市场这样复杂的非线性特征系统，人工神经网络可以达到更加准确的预测精度，已经成为了股票价格预测方法的研究热点。

[0012] 目前，一些基于深度学习的模型已经被应用到股票市场分析方面，Dixon等人使用深度神经网络预测了43种商品和期货在5分钟内的价格变化，使用随机梯度下降的BP算法进行训练，实现了42％的准确率。Fehrer和Feuerriegel构建了一个基于新闻头条的德国股票收益模型。他们使用递归自编码器，在每一个自编码器上都有一个用于估计概率的附加softmax层，分成三类{-1,0,1}用来预测下一天与新闻头条相关的股票收益。该方法首先使用高斯噪声(Gaussian noise)对权重进行初始化，然后通过反向传播进行更新，结果显示递归自编码器达到56％的准确度。Xiong等人通过对开盘价走势和收盘价估计每日标准普尔500的日常波动，采用LSTM模型将每日标准普尔500的收益率、波动率以及25个国内主要领域的谷歌趋势作为输入，并采用平均绝对百分误差(MAPE)作为目标损失函数.结果显示，其优于其他对比模型约31％。

[0013] 有效市场假说指出，证券价格反映了所有可用的信息，每个人都有一定程度的信息获取。研究人员认为，可以通过比较历史价格和成交量波动率与当前价格的关系来进行预测。但这样的研究方法忽略了股票市场价格波动的关键来源——财经新闻。利用多种信息源能够获得比单一信息更高的预测精度，而自然语言处理技术的进步为研究财经新闻对股票市场价格波动的影响提供了可能，但是若直接简单地将所有新闻的影响加入，模型可能会受到噪声的干扰。同时，新闻文本篇幅较长，生成的矩阵维数过高，极易造成维数灾难。

发明内容

[0014] 为了克服上述现有技术中的不足，本发明提供一种结合财经新闻的卷积神经网络股票价格波动预测方法，方法包括：

[0015] 步骤一：扫描语料库，预设关键词，设定扫描窗口长度，在设定的扫描窗口内配置共现矩阵；

[0016] 步骤二：基于共现矩阵配置共现概率；

[0017] 步骤三：将语料库中词语转化为索引矩阵，设定要保留的关键词个数，设定模型参数，基于共现概率定义词向量矩阵；将数据集中的数据切分成训练集和测试集两部分，利用索引与训练出来的关键词矩阵一一对应，获得训练集的词向量；

[0018] 步骤四：基于预测股价的涨跌作为一个分类问题，即涨和跌两类；将训练集和测试集的词向量矩阵载入卷积神经网络，将新闻词向量和股票利用日期信息一一对应，利用股价的涨跌作为特征标签，将训练集和测试集分别提取出特征标签和类别标签；

[0019] 步骤五：预测模型的网络结构为5层，其中前两层为卷积层，第一层神经元个数为64个，卷积核大小为3*100，输入矩阵大小30*100，激活函数为ReLU函数，第二层卷积神经元个数为32，卷积核大小为3*50，第三层为最大池化层，池化大小为18*1；

[0020] 利用18*1的窗口扫过矩阵，提取每个窗口中的最大值；得到一个新矩阵；

[0021] 第四层和第五层为全连接层，第四层的全连接层神经单元节点数为64，激活函数为ReLU函数；

[0022] 第五层激活函数则是softmax，是进行最后的分类，得到了每只股票后一天股价的涨跌情况。

[0023] 优选地，步骤一还包括：

[0024] 设定窗口长度为n；对设定窗口中的句子进行扫描，得到关键词i，j在设定窗口中出现的次数X_{ij}.遍历整个语料库后得到共现矩阵X。

[0025] 优选地，步骤二还包括：计算关键词i，j的共现概率

[0026] $$P_{ij}=P(j|i)=\frac{X_{ij}}{X_i}, \quad (1)$$

[0027] 这个比率反映了词语之间的相关性，称词i和j分别为中心词和背景词，接下来是利用模型训练词向量，使用词向量表达共现概率的比值，任意一个比值需要三个单词i，j，k：

[0028] $$F(w_i,w_j,w_k)=\frac{P_{ik}}{P_{jk}}, \quad (2)$$

[0029] 这里，w是单词的词向量，P_{ik}是单词k出现在单词i上下文的概率，P_{jk}是单词k出现在单词j上下文的概率.$F(w_i,w_j,w_k)$是关于词向量的函数，由于向量的本质是线性结构，式(2)中F的形式是F＝exp，因此可以推出词向量需要满足的等式：

[0030] $$\log(X_{i,j})=v_i^T v_j+b_i+b_j, \quad (3)$$

[0031] 损失函数为：

[0032] $$J=\sum_{i,j=1}^{V} f(x_{ij})(\omega_i^T\omega_j+b_i+b_j-\log x_{i,j})^2, \quad (4)$$

[0033] V是词汇量的大小，f(x)是权重函数，b_i和b_j是偏置项，权重函数f(x)可以参数化为：

[0034] $$f(x)=\begin{cases}(x/x_{\max})^{\alpha} & \text{if } x<x_{\max}\\ 1 & \text{otherwise}\end{cases}, \quad (5)$$

[0035] α等于$\frac{3}{4}$；

[0036] 得到全部词向量之后，每个词向量都是中心词和背景词向量之和。

[0037] 优选地，步骤三中：预设保留的关键词个数为3000至4000个。

[0038] 从以上技术方案可以看出，本发明具有以下优点：

[0039] 本发明利用自然语言处理技术来提取相关新闻中的特征，从而分析、观测财经新闻与股票价格走势的关联程度。本文结合新闻报道的有效信息，提出了一种基于卷积神经网络的股票价格波动预测方法。首先，将新闻分词，并提取主要事件，利用出现次数最多的前3000个词语作为关键词，并使用Glove模型将其表示为低维稠密的词向量；其次，将新闻特征和股票价格对应起来，把时间划分成短、中、长三个时间段，用卷积神经网络来模拟新闻事件对股票价格变动的短期和长期影响；最后，通过训练好的模型预测股票的涨跌情况。

[0040] 本发明将自然语言处理和卷积神经网络技术相结合，应用到股票市场分析、预测中，将新闻表示为词向量的形式，从中提取特征在卷积神经网络中训练，实验结果证明，利用卷积神经网络输出分类结果的正确率高于传统方法。

附图说明

[0041] 为了更清楚地说明本发明的技术方案，下面将对描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0042] 图1为结合财经新闻的卷积神经网络股票价格波动预测方法流程图；

[0043] 图2为训练词向量流程图；

[0044] 图3为基于CNN神经网络的预测模型结构图；

[0045] 图4为卷积层数对模型的影响示意图；

[0046] 图5为模型精确率与MCC对比示意图；

[0047] 图6为迭代次数对实验结果的影响示意图。

具体实施方式

[0048] 本发明提供一种结合财经新闻的卷积神经网络股票价格波动预测方法，如图1所示，方法包括：

[0049] S1：扫描语料库，预设关键词，设定扫描窗口长度，在设定的扫描窗口内配置共现矩阵；

[0050] 假设关联度高的词语出现在相同文档的可能性高，所以每个单词都可以用周边的词来表示。设定窗口长度为n；对设定窗口中的句子进行扫描，得到关键词i，j在设定窗口中出现的次数X_{ij}.遍历整个语料库后得到共现矩阵X。

[0051] S2：基于共现矩阵配置共现概率；

[0052] 计算关键词i，j的共现概率

[0053] $$P_{ij}=P(j|i)=\frac{X_{ij}}{X_i}, \qquad (1)$$

[0054] 这个比率反映了词语之间的相关性，称词i和j分别为中心词和背景词，接下来是利用模型训练词向量，使用词向量表达共现概率的比值，任意一个比值需要三个单词i，j，k：

[0055] $$F(w_i,w_j,w_k)=\frac{P_{ik}}{P_{jk}}, \qquad (2)$$

[0056] 这里，w是单词的词向量，P_{ik}是单词k出现在单词i上下文的概率，P_{jk}是单词k出现在单词j上下文的概率.$F(w_i,w_j,w_k)$是关于词向量的函数，由于向量的本质是线性结构，若考虑相似关系，可以利用向量的差.而单词的共现矩阵、待预测的单词k和上下文单词i，j之间的区别是任意的，也就是说，它们之间的关系是对称的，可以交换这两种词之间的位置，而模型在这种调换之下的意义应该是不变的.式(2)中F的形式是F＝exp，因此可以推出词向量需要满足的等式：

[0057] $$\log(X_{i,j})=v_i^T v_j+b_i+b_j, \qquad (3)$$

[0058] 损失函数为：

[0059] $$J=\sum_{i,j=1}^{V} f(x_{ij})(\omega_i^T\omega_j+b_i+b_j-\log x_{i,j})^2, \quad (4)$$

[0060] V是词汇量的大小，f(x)是权重函数，b_i和b_j是偏置项，权重函数f(x)可以参数化为：

[0061] $$f(x)=\begin{cases}(x/x_{max})^{\alpha} & \text{if } x<x_{max}\\ 1 & \text{otherwise}\end{cases}, \quad (5)$$

[0062] α等于$\frac{3}{4}$；

[0063] 得到全部词向量之后，每个词向量都是中心词和背景词向量之和。Glove模型结合了LSA和word2vec模型的优点，既利用了语料库的全局信息，又利用了上下文局部信息，其训练速度更快，对于大规模语料算法和小语料库的扩展性表现都很好。

[0064] S3：将语料库中词语转化为索引矩阵，设定要保留的关键词个数，设定模型参数，基于共现概率定义词向量矩阵；将数据集中的数据切分成训练集和测试集两部分，利用索引与训练出来的关键词矩阵一一对应，获得训练集的词向量；

[0065] 预设保留的关键词个数为3000至4000个。

[0066] S4：基于预测股价的涨跌作为一个分类问题，即涨和跌两类；将训练集和测试集的词向量矩阵载入卷积神经网络，将新闻词向量和股票利用日期信息一一对应，利用股价的涨跌作为特征标签，将训练集和测试集分别提取出特征标签和类别标签；

[0067] S5：预测模型的网络结构为5层，其中前两层为卷积层，第一层神经元个数为64个，卷积核大小为3*100，输入矩阵大小30*100，激活函数为ReLU函数，第二层卷积神经元个数为32，卷积核大小为3*50，第三层为最大池化层，池化大小为18*1；

[0068] 利用18*1的窗口扫过矩阵，提取每个窗口中的最大值；得到一个新矩阵；这一层作用是降维同时提取出最主要的特征。

[0069] 第四层和第五层为全连接层，第四层的全连接层神经单元节点数为64，激活函数为ReLU函数；

[0070] 第五层激活函数则是softmax，是进行最后的分类，得到了每只股票后一天股价的涨跌情况。

[0071] 本发明爬取了沪深上市公司中前500名的股票信息，再利用新浪财经的接口tushare获取了2016-01-01至2018-3-14之间的所有新闻信息，包括股票代码、日期、新闻标题和内容的链接。

[0072] 本发明的目标是从新闻中提取出简洁有用的信息来反映股市的变化。本发明仅从新闻标题中提取事件，利用中文分词模块，将新闻标题分成单词，例如：“关于股东股票解除质押的公告”，可以分为“关于/股东/股票/解除/质押/的/公告”。所有新闻都拆分成类似的结构之后，将每个词对应到训练好的词向量上。

[0073] 早期自然语言处理的方法是将单词表示为原子符号，缺点是工作太过繁琐、效率低下.而词向量技术是将词转化成稠密向量，对于相似的词，其对应的词向量也相近.在自然语言处理应用中，词向量作为深度学习模型的特征进行输入.最终模型的效果很大程度上取决于词向量的效果.语言的语义向量空间模型用实值向量表示每个单词，这些载体可以作为各种信息检索、文档分类、命名实体识别等。

[0074] 词向量最初训练方法是one-hot编码，将词库中每个单词编码成相同维度的向量，每个词在对应词库位置的维度编码为1，其余维度是0.这样的编码方式忽视了词与词之间的关系，形成的词向量不能表示单词的相似性.而基于窗口的共现矩阵方法基本思想是任一词的含义可以用它的周边词来表示，在低维度的情况下保留了更多的句子信息，即词和词的相似性。

[0075] 基于词共现学习单词向量的两个主要模型为：(1)全局矩阵分解方法，如潜在语义分析(LSA)；(2)局部上下文窗口方法，Mikolov提出了word2vec模型。LSA的基本思想是利用了矩阵分解，将文档表示成行，单词表示为列，第i行第j列表示的是文档i中是否包含单词j。通过SVD矩阵分解的方式得到两组向量分别表示文档的向量和单词的向量。其优点在于利用了词共现的信息，不仅仅关注窗口大小的文档信息.但在单词类比任务上表现得相对较差.word2vec模型通过预测单词周边其他单词出现次数来学习低维度的词向量.基本思想是比较单词在文档出现的上下文环境，例如单词“餐厅”和“饭馆”在“我吃饭”这种环境的上下文中出现次数很高，那么可以认为单词”餐厅“和”饭馆“的词向量就比较相似.但是该方法对每个上下文窗口单独训练，没有利用包含在共现矩阵中的统计信息.因为使用了唯一词向量，对多义词处理乏力。

[0076] 基于分析产生线性意义方向所需的模型属性前提下，采用了一个特定的加权最小二乘模型Glove，它结合了上述两种方法的优点，通过对单词共现矩阵中的非零元素进行训练来有效利用统计信息.产生具有意义的子结构的向量空间.训练词向量过程如图2所示。

[0077] 共现矩阵的基本思想是：假设关联度高的词语出现在相同文档的可能性高，所以每个单词都可以用周边的词来表示.首先，设定窗口长度为n的统计窗口对语料库中的句子进行扫描，得到单词i，j在上下文中出现的次数X_{ij}.遍历整个语料库后得到共现矩阵X.计算每对词的共现概率

[0078] $$P_{ij}=P(j|i)=\frac{X_{ij}}{X_i}, \qquad (1)$$

[0079] 这个比率反映了词语之间的相关性.称词i和j分别为中心词和背景词.接下来是利用模型训练词向量，使用词向量表达共现概率的比值.任意一个比值需要三个单词i，j，k：

[0080] $$F(w_i,w_j,w_k)=\frac{P_{ik}}{P_{jk}}, \qquad (2)$$

[0081] 这里，w是单词的词向量，P_{ik}是单词k出现在单词i上下文的概率，P_{jk}是单词k出现在单词j上下文的概率.$F(w_i,w_j,w_k)$是关于词向量的函数.由于向量的本质是线性结构，若考虑相似关系，可以利用向量的差.而单词的共现矩阵、待预测的单词k和上下文单词i，j之间的区别是任意的，也就是说，它们之间的关系是对称的，可以交换这两种词之间的位置，而模型在这种调换之下的意义应该是不变的.式(2)中F的形式是F＝exp，因此可以推出词向量需要满足的等式：

[0082] $$\log(X_{i,j})=v_i^T v_j+b_i+b_j, \qquad (3)$$

[0083] 损失函数为：

[0084] $$J=\sum_{i,j=1}^{V} f(x_{ij})(\omega_i^T\omega_j+b_i+b_j-\log x_{i,j})^2, \quad (4)$$

[0085] V是词汇量的大小，f(x)是权重函数，b_i和b_j是偏置项.权重函数f(x)可以参数化为：

[0086] $$f(x)=\begin{cases}(x/x_{\max})^{\alpha} & \text{if } x<x_{\max}\\ 1 & \text{otherwise}\end{cases}, \quad (5)$$

[0087] 根据经验，当α等于$\frac{3}{4}$时效果最好.

[0088] 最后，得到全部词向量之后，每个词向量都是中心词和背景词向量之和.Glove模型结合了LSA和word2vec模型的优点，既利用了语料库的全局信息，又利用了上下文局部信息(滑动窗口)，其训练速度更快，对于大规模语料算法和小语料库的扩展性表现都很好.

[0089] 基于上述理论，本发明选择用Glove模型训练新闻标题文本.为了使训练到的词向量尽可能多的匹配新闻数据，将沪深股市2010年至2016年期间前500支股票新闻作为语料库进行训练.首先将语料库中每个句子进行分词，排除掉逗号、句号、冒号等中文符号，统计所有单词出现的次数，设定要保留的关键词个数为3000个，其余均设为标志“unkonwn”.遍历完整个语料库之后得到一个关键词矩阵，采用”gb18030”方式编码.之后，建立Glove模型，本发明设定模型的每个单词维数为50，每条新闻最大用25个单词表示，窗口大小10.学习率为10^{-5}，权重函数中的α为0.75.将关键词矩阵和单词维数以及新闻作为变量输入Glove模型中，最终得到了训练好的词向量矩阵。

[0090] 本发明中，新闻和股票数据的预处理过程以及模型的构建过程，利用tushare接口中提供的函数首先获取到了沪深股市上市公司前500名的公司股票数据，包括公司股票代码、名称、流通股本以及总资本等.得到股票代码列表之后，首先去掉没有新闻的股票，得到了所有可用的股票列表和对应的财经新闻.然后直接调用接口函数获取列表中股票的前复权数据，包括股票代码、当日开盘价、当日收盘价、当日最高阶和最低价以及成交总额.对于股票预测来说，复权功能可以消除由于除权除息造成价格指标走势畸变的现象，判断当前股价是否处于相对历史高位还是低位.由于需要利用历史价格来预测股价的涨跌，所以选择前复权数据进行预测.由于只预测涨跌，不涉及具体变化幅度，本发明选择的对比基准是沪深300指数，通过计算相对回报率，将每只股票与沪深300的前复权收盘价取对数再相减，获得每只股票相对于沪深300指数的涨跌情况。

[0091] 将长期事件看作过去一个月的事件，中期事件作为过去一周的事件，短期事件为过去一天的股票涨跌变化.可知不同时间跨度的事件对股票的影响效果不同.基于CNN的预测模型学习了这三个时间跨度数据对股票价格涨跌的影响，其网络结构图如图3所示。

[0092] 模型的输入是训练好的新闻词向量矩阵，事件按照时间长短排序，把每只股票最后一百天的价格和新闻数据作为测试集，其他均为训练集.模型的输出分为两类，类别0代表股票价格下跌，类别1代表股票价格增加.本发明在CNN网络结构上做了细微的改变，设k是句子中第i个单词对应的单词向量维数，长度为n的句子可以表示为(长度不足时用零向量来填充)：

[0093] $$x_{1:n}=x_1\oplus x_2\oplus\mathrm{L}\oplus x_n, \quad (6)$$

[0094] ⊕为连接运算符，令$x_{i:i+j}$作为单词x_i, x_{i+1}, L, x_{i+j}的连接。

[0095] 模型的前两层是卷积层，且每一层卷积后加入一层激活层，激活函数采用ReLU，卷积运算包含一个滤波器$\omega \in R^{hk}$，它应用于h个单词的窗口来提取新的特征.例如，特征c_i是由$X_{i,i+h-1}$的窗口产生：

[0096] $$c_i = f(\omega * x_{i:i+h-1} + b), \quad (7)$$

[0097] 其中*表示卷积操作，$b \in R$是偏置项，f是双曲线正切的非线性函数。该滤波器应用于句子$\{x_{1:h}, x_{2:h+1}, L, x_{n-h+1:n}\}$每个可能的单词窗口来产生特征映射$c = [c_1, c_2, L, c_{n-h+1}]$，这个操作即为基于滑动窗口的特征提取。

[0098] 为了利用股票的本地特征和全局特征来预测价格的走势，最后一层卷积层的顶部加入了一个最大池化层，使神经网络只保留由卷积层产生的最有用的本地特征。

[0099] 给定一系列新闻事件输入序列$U = (U_1, U_2, L, U_n)$，$U_i \in R^d$，一维卷积函数取新闻输入序列U中每个事件的权重向量$w_1 \in R^1$，得到一个新序列Q：

[0100] $$Q_j = w_i^T U_{j-l+1:j}, \quad (8)$$

[0101] 为了确定全局特征中最具有代表性的特征，本发明在Q上执行最大池化操作$V_j = \max Q(j:\bullet)$，其中$Q(j:\bullet)$是矩阵Q的第j行，在最大池化操作后得到特征向量v，对于长期和中期时间，分别得到特征向量V和V_m，而对于短期事件，特征向量V_s是直接对短期事件序列取平均值.最终得到的特征向量是长期序列、中期序列及短期序列组合起来的特征向量$V = (V_1, V_m, V_s)$。

[0102] 为了防止过拟合，每一层神经网络都随机让一部分神经元失去功能.令Y为池化层的输出，全连接层操作函数为：

[0103] $$y_{cls} = f(Y) = \sigma\left(w_1^T \cdot Y\right), \quad (9)$$

[0104] σ为softmax函数，w_1^T是权重向量.得到输出：

[0105] $$y_{cls}(cls \in \{0,1\}), \quad (10)$$

[0106] cls为网络输出值.迭代次数为十次，将每次模型训练出的结果表示成混淆矩阵，用于计算精确率和相关系数.在机器学习中，混淆矩阵是评价分类模型的形象化工具，矩阵的每一列表示模型预测的样本情况，矩阵的每一行表示样本的真实情况。

[0107] 为了进一步说明本发明涉及的方案，下面以具体的实施例予以详细说明。

[0108] 本发明利用新浪财经数据接口Tushare获取了沪深上市公司中排名前500公司的信息地雷数据和股票价格数据，时间范围是2016-01-01至2018-03-16.其中，这里基于新闻标题做出的预测。本发明选择从新闻标题中选取事件，对沪深300指数和沪深个股进行了预测.实验分三个时间间隔进行：短期(1天)、中期(7天)、长期(28天).利用图3训练的网络结构模型来预测对于不同时长间隔的股票价格波动.

[0109] 将91500个交易日数据按时间间隔分为短中长三部分，分别与对应前一天、前一周和前一个月的新闻数据对齐.新闻数据划分为训练集、验证集和测试集，其中测试集为每个时间间隔的最后一百天数据，占1/6，验证集占总数据量的1/6，剩下的2/3用于训练.如表1所示.

[0110] 表1数据集分布

[0111]

	Short	*Mid*	*Long*
The number of Train	15000	13000	13500
The number of Validation	3400	3250	3375
The number of Test	3512	3330	3000

[0112] 评价指标

[0113] 上文提到基于预测值和真实值的关系，模型的预测结果以混淆矩阵的形式表示，可以将样本分为四个部分，分别是：

[0114] 真正例(True Positive,TP)：预测值和真实值都为1.

[0115] 假正例(False Positive,FP)：预测值为1，真实值为0.

[0116] 真负例(True Negative,TN)：预测值与真实值都为0.

[0117] 假负例(False Negative,FN)：预测值为0，真实值为1.混淆矩阵的表示如下：

[0118] 表2混淆矩阵

[0119]

	Positive	*Negative*
True	TP	TN
False	FP	FN

[0120] 本发明使用的评价指标是基于混淆矩阵计算的。第一个是准确率，这是最直观的衡量预测结果的指标.但是精确率对数据分布非常敏感，当预测股票涨或跌其中一种情况分布特别多时，使用对大对数类别进行预测的分类器，精确率会很高，对于模型的评价不够客观。

[0121] $$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \qquad (11)$$

[0122] 在之前的工作中，使用马修斯相关系数(MCC)来避免由于数据倾斜造成的偏差.马修斯相关系数是一个单一的汇总值，包含了混淆矩阵的所有单元格.衡量模型性能的常用指标，本质上是预测结果和观察结果之间的相关性。

[0123] $$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \qquad (12)$$

[0124] 新闻对预测结果的影响

[0125] 本发明的预测模型是使用Glove模型来学习词向量，从新闻文本中提取最能代表新闻特点的关键词且最大程度的保留词和词之间的语义联系，并建立了基于卷积神经网络的预测模型.为了展示新闻作为影响因素的重要性，本发明分别将新闻词向量+股票价格和只有股票价格两种情况输入模型中，观察预测准确率.对比结果如表3所示，新闻词向量有效提高了准确率及MCC的结果。

[0126] 表3有无新闻实验结果对比

[0127]

	Accuracy	*MCC*
News+price	65.97%	0.437
Only price	42.02%	0.023

[0128] 结果显示，加入新闻之后模型的预测准确率大大提升，若影响因素只有股票价格，准确率与MCC值变得很低.因此，新闻事件内容对股票价格的波动影响重大。

[0129] 不同卷积层数对预测结果的影响

[0130] 模型的激活函数是ReLU函数，这里采用不同数量的卷积层进行对比，结果如表4所示.两个卷积层的精确率略优于一个卷积层，说明多卷积层可以解释更复杂的分类关系。但是，层数越多，训练速度越慢，所以在实验分析中只分析卷积层数从1层到3层的预测情况。

[0131] 表4不同数据量对实验结果的影响

[0132]

		Short	*Mid*	*Long*	*Traing speed*
1-layer	Accuracy	64.71%	60.59%	62.71%	1s
	MCC	0.33	0.21	0.27	
2-layer	Accuracy	65.97%	61.68%	64.19%	43s
	MCC	0.43	0.25	0.26	
3-layer	Accuracy	60.31%	58.73%	58.83	83s
	MCC	0.19	0.16	0.16	

[0133] 只观察短期事件影响，可以很清晰观察出卷积层数对模型影响.结果如图4所示。

[0134] 结果表明，两层卷积的结果最好，当模型达到三层卷积的时候准确率明显下降，速度上越来越慢.因此，考虑综合因素，两层卷积的神经网络模型的性价比最好.其次，短期事件的预测结果优于中期事件和长期事件.原因可能是价格反应和新闻信息之间存在一天的延迟。也存在某些事件一经出现立即导致股价发生变化，

[0135] 新闻数据量对预测结果的影响

[0136] 本发明从新浪财经网站分别提取新闻标题和内容进行了实验。虽然标题可以提供关于新闻的中心信息，但内容可以提供一些背景知识或细节。

[0137] 本发明主要使用标题，本发明设计一个比较实验来分析新闻标题和内容的有效性。使用新浪财经新闻来比较新闻标题和内容的有效性，然后添加新闻标题来调查数据量是否重要.表5显示只使用新闻标题可以获得最佳表现.分析认为，可能的原因是会从新闻内容中提取一些不相关的事件干扰预测结果。

[0138] 表5不同数据量对实验结果的影响

[0139]

	Title	*Content*	*Content+title*
Accuracy	65.97%	60.33%	64.07%

[0140]

MCC	0.43	0.19	0.26

[0141] 对比实验

[0142] 为了进一部分析模型的性能，将本发明方法与已有的模型进行对比。

[0143] 1.Luss and Aspremont:Luss和Aspremont在2012年提出来的，利用支持向量机(SVM)构造预测模型。SVM是一种线性分类模型，训练集包括新闻文档和输出类别，特征由词袋模型确定，通过线性函数确定分类类别.是一种比较先进的基于新闻的股市预测模型。

[0144] 2.WB-NN:同样使用词向量，使用标准前馈神经网络(NN)来建立模型，用于与卷积神经网络进行对比。

[0145] 3.E-NN:根据丁效在2014年提出来的采用结构化事件元组$E=(O_1;P;O_2)$来代表新闻文档，通过标准前馈神经网络来研究事件和股票价格变动之间的关系。

[0146] 本发明方法与以上方法的对比结果如表6和图5所示.

[0147] 表6实验方法结果对比

[0148]

	Accuracy	*MCC*
Luss and Aspremont	56.42%	0.071
WB-NN	60.25%	0.195
E-NN	62.84%	0.347
Our model	65.97%	0.433

[0149] 考虑到迭代次数对实验结果的影响，实验对E-NN、WB-NN和本发明模型进行10次迭代，最终实验结果取最好的一次迭代结果。迭代次数对实验结果的影响如图6所示.可以观察到，模型准确率随迭代次数增加而增大，但到达一定值后开始下降，其中WB-NN和E-NN都是在迭代次数7次以后达到最大，而本发明模型在5次左右准确率达到最大并在此后开始下降。

[0150] 本发明又测试了卷积神经网络与前馈神经网络模型、基于词向量和以单词表示文档以及利用结构化事件元组代表文档三种方法.得出结论如下：

[0151] 1)与上述对比方法相比，本发明模型均取得了更好的成绩.就综合因素而言，只考虑短期事件的影响，模型为2层卷积，第一层取64个神经元、第二层32个神经元、阈值＝0.6时，预测效果最好，准确率可以达到65.974％.

[0152] 2)卷积神经网络模型的性能优于基于SVM的预测模型，在卷积神经网络中可以学习新闻事件与股票价格之间的隐藏关系，同时在Luss的实验里，词袋模型并没有结合新闻中词语之间的联系，所以提取到的特征并没有词向量准确.

[0153] 3)基于关键词的词向量的比只利用大量单词来代表文档的准确性更高.可能的原因有以下几点：首先，低维向量可以有效的解决特征稀疏问题.其次，关键词可以最大程度的代表新闻的含义，如果不对新闻文本做一个筛选，其中某些单词可能成为噪音，干扰新闻原本的语义情感.

[0154] 4)在同样使用词向量的情况下，基于卷积神经网络的模型优于前向神经网络的模型(WB-NN)，因为卷积神经网络可以定量分析更长历史事件的影响，最主要的原因是在预测过程中，通过卷积操作提取更具有代表性的特征向量，而前馈神经网络的自身和层与层之间没有连接，所以提取到的特征没有卷积神经网络准确.

[0155] 新闻的质量比数量更重要.也就是说，最相关的信息(例如新闻标题)比更多但相关度更低的信息要好.

[0156] 本发明通过分析股票价格和所对应的财经新闻在不同长度的时期对股票波动情况形成的综合影响，本发明得到相应的实验结果，基于新闻词向量的表示方法优于离散事件，卷积神经网络可以捕捉新闻事件的长期影响，优于标准前馈神经网络。

[0157] 除了财经新闻之外，情感是新闻文档语义分析的另一个视角.Tetlock研究如何将定性信息(即特定新闻专栏中负面词语的分数)纳入总体市场估值中。Si等人建议回归主题情感时间序列和股票价格时间序列。他们的工作与股票市场预测是正交的。财经新闻中，有一部分词语带有很强烈的正面或负面影响，在接下来的工作中，本发明考虑将情感分析纳入股票市场预测中，挖掘词语的深层含义，同时结合投资策略优化模型具备盈利能力。

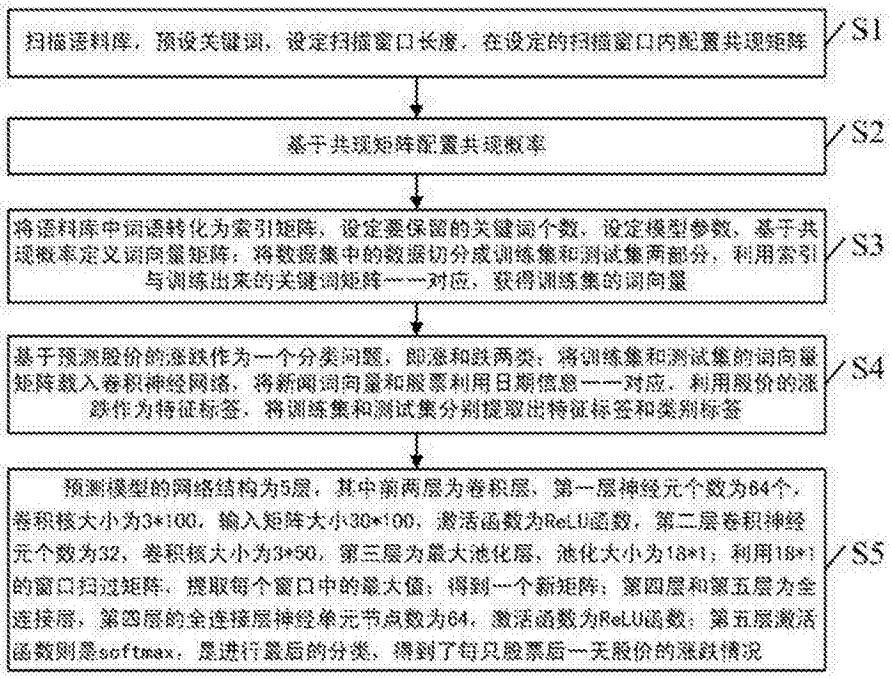
扫描语料库，预设关键词，设定扫描窗口长度，在设定的扫描窗口内配置共现矩阵
S1
基于共现矩阵配置共现概率
S2
将语料库中词语转化为索引矩阵，设定要保留的关键词个数，设定模型参数，基于共现概率定义词向量矩阵，将数据集中的数据切分成训练集和测试集两部分，利用索引与训练出来的关键词矩阵一一对应，获得训练集的词向量
S3
基于预测股价的涨跌作为一个分类问题，即涨和跌两类，将训练集和测试集的词向量矩阵载入卷积神经网络，将新闻词向量和股票利用日期信息一一对应，利用股价的涨跌作为特征标签，将训练集和测试集分别提取出特征标签和类别标签
S4
预测模型的网络结构为5层，其中前两层为卷积层，第一层神经元个数为64个，卷积核大小为3*100，输入矩阵大小30*100，激活函数为ReLU函数，第二层卷积神经元个数为32，卷积核大小为3*50，第三层为最大池化层，池化大小为18*1，利用18*1的窗口扫过矩阵，提取每个窗口中的最大值，得到一个新矩阵，第四层和第五层为全连接层，第四层的全连接层神经单元节点数为64，激活函数为ReLU函数，第五层激活函数则是softmax，是进行最后的分类，得到了每只股票后一天股价的涨跌情况
S5

图1

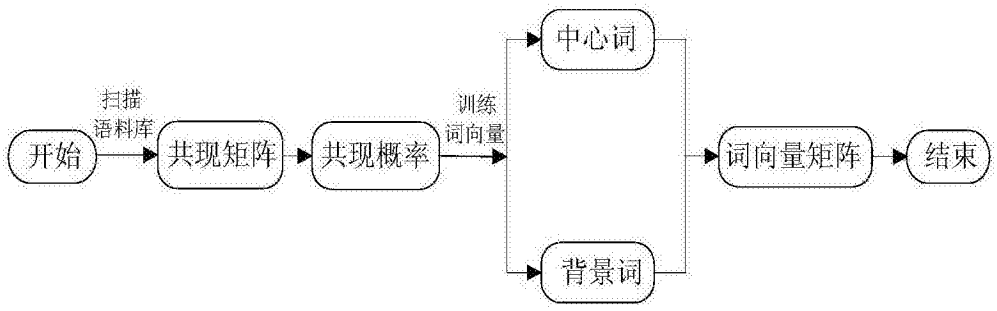
开始
扫描语料库
共现矩阵
共现概率
训练词向量
中心词
背景词
词向量矩阵
结束

图2

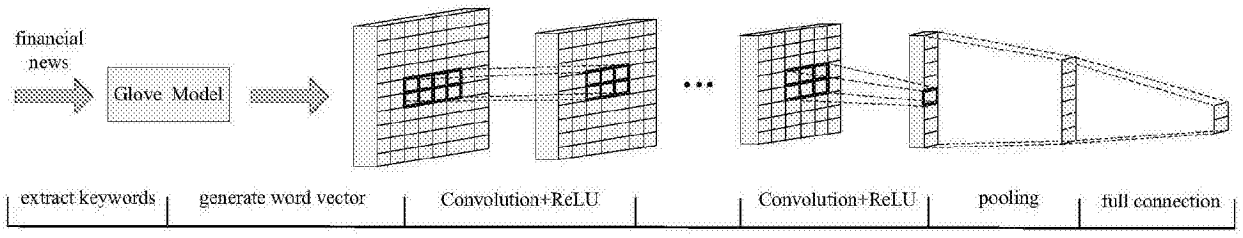
financial news
Glove Model
extract keywords
generate word vector
Convolution+ReLU
Convolution+ReLU
pooling
full connection

图3

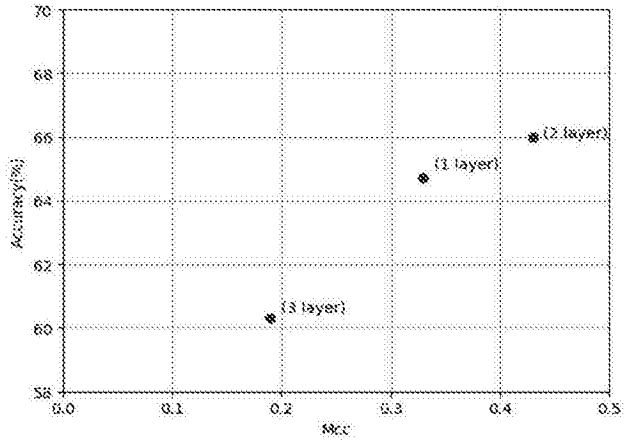
70
68
66
64
62
60
58
Accuracy(%)
0.0
0.1
0.2
0.3
0.4
0.5
Mcc
(2 layer)
(1 layer)
(3 layer)

图4

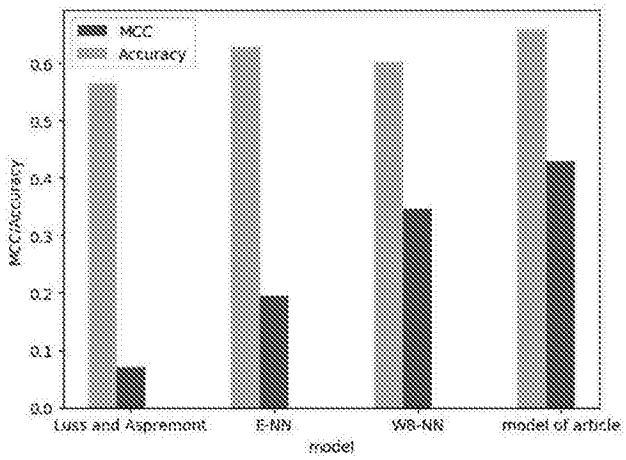
MCC
Accuracy
0.6
0.5
0.4
0.3
0.2
0.1
0.0
MCC/Accuracy
Loss and Aspremont
E-NN
WB-NN
model of article
model

图5

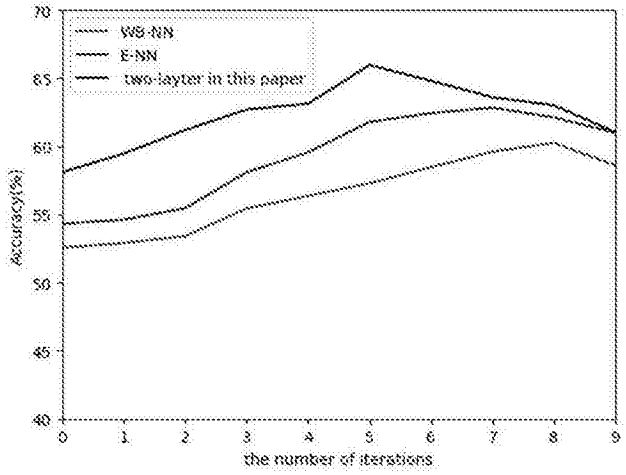
WB-NN
E-NN
two-layter in this paper
70
65
60
55
50
45
40
Accuracy(%)
0
1
2
3
4
5
6
7
8
9
the number of iterations

图6