# 正文

## 问题描述

#### 社交网络

作为实时在线算法，首先考虑时间因素，定义进行计算的时间窗口为

定义当前时刻社交网络语料库中所有的词构成集合，所有的句子构成集合.

定义社交媒体上舆论讨论的对象的集合为：

定义参与讨论的用户为：

#### 情感评价

根据当前时刻社交网络语料中的所有句子以及词汇信息，需要找到来自于参与讨论用户对特定讨论对象情感评价结果。

定义用户对于对象的情感评价为：

其中，

为用户对于对象给出的情感评价。

考虑对于对象，当前时刻现有所有的对于该对象的语言评价到情感得分的矩阵记为：

即对象的情感得分为：

#### 股价变动

根据当前时刻特定讨论对象的情感评价结果，需要找到情感评价结果对相应上市公司下一时刻的股价的影响关系。定义对象对应上市公司股价的变动为

其中为行业轮动等原因造成的当前时间窗口股价对下一时间窗口股价的影响结果，为情感得分对股价的影响结果。

#### 性能指标

对情感分析的性能衡量主要考虑其准确率，AUC，以及F1分数。

假设分类结果的真阳性，假阳性，真阴性，假阴性分别为：. 其计算方法分别为：

AUC由ROC曲线得到

对股价预测主要考虑和真实股价相比的平均绝对百分比误差以及预测方差。

考虑时间窗口q时刻的预测得分以及平均绝对百分比误差：

预测方差为

## 算法设计

#### 自然语言模型化

###### Word2vec算法

由于在基于上下文的情感分析领域中word2vec算法的效果较好，于是考虑使用Skip-Gram与CBOW两种word2vec模型[17]进行模型化尝试。两种模型的结构都是需要利用神经网络的结构。所以word2vec模型在训练时间上的消耗会明显增多，但预期的训练效果优于模型简单的PCA/TF-IDF算法。

CBOW模型的训练输入是某一词汇上下文的词汇的词向量，输出是该词汇的词向量，即根据上下文的词汇推断该处的词汇。Skip-Gram模型的训练输入是某一词汇的词向量，输出是给定词汇上下文的词向量，即根据该处词汇推断上下文的词汇。Word2vec使用霍夫曼树代替传统神经网络中的隐藏层和输出层的神经元，其叶子结点起到输出层的作用，内部结点起到隐藏层的作用。输入层到隐藏层与传统神经网络采用线性变换加激活函数不同，而是对所有的输入词向量求和并求平均。隐藏层到输出层为了不计算所有词的softmax概率，采用霍夫曼树表示隐藏层与输出层。使用二元逻辑回归的方法，将霍夫曼树左子树定义为负类，右子树定义为正类，一般使用sigmoid函数进行判别：

在式（2-5）中，是当前内部结点的词向量，是需要训练出来的模型参数。

###### Skip-Gram模型

Skip-Gram模型输入为文本库中某个词的词向量，训练对象为该词汇词上下文的词向量[17]。Skip-Gram模型定义如下：

定义的某一子集构成中的给定元素上下文的词集合，集合大小为;

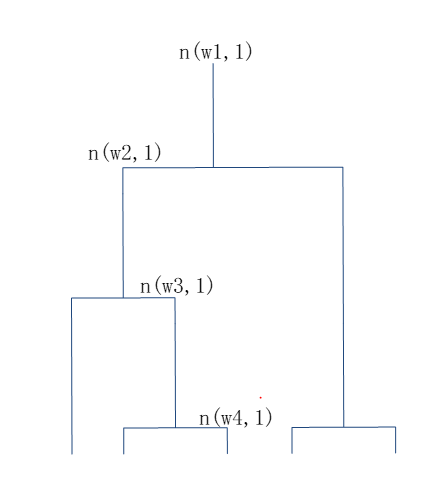
定义给定对应的数据样本为;

定义输出层包含样本中词向量;

定义投影层为恒等投影.

然后以各词在语料中出现过的词当作叶子结点，以所有词w在语料中出现的频率作为权值构建霍夫曼树（Huffman Tree）. 满足叶子结点个，其他结点个。

例如，在图2-1所示的霍夫曼树的结构中：



以计算结点为例，计算的过程，实际上是期望最大化函数：

通过最大化所有节点的似然函数乘积，便可以得到最后的迭代结果。具体做法是，每次仅用一个样本更新梯度，减少梯度计算量。得到的对数似然函数是：

由此得到梯度表达式为：

通过梯度上升法，进一步迭代求出和。

CBOW模型与Skip-Gram类似，篇幅所限不再赘述

###### Hierarchical Softmax

针对Skip-Gram与CBOW模型，采用Hierarchical Softmax的方法对问题进行简化[17]。在Huffman树中，考虑某个出现过的词*w*对应的叶子结点：

1. 定义为根节点出发到达的路径；

2. 定义为路径的结点个数；

3. 定义为路径的所有结点；

4. 定义为词在中的Huffman编码；

5. 定义为中非叶子结点对应的向量。

使用sigmoid函数作为激活函数，其词汇条件概率的 为:

其中：

其对数似然函数为

记

采用随机梯度上升法，对于的梯度为：

由此可以得到迭代公式为：

其中，表示学习率。

对于的梯度计算为：

的迭代公式为：

在Skip-Gram模型中，推导相类似。

#### 在线SVM算法情感分类

定义时间窗口内新到来的的某一条/多条评价矩阵记为：

为内所有新到来的自然语言模型数据。

首先不考虑新到来的模型数据，传统的基于软边界的SVM算法是找到能分割训练数据的核的最佳函数的线性组合：

一般考虑最小化下面的目标函数：

一般使用拉格朗日对偶的KT条件求解，即满足：

现在考虑边界向量会随着新数据加入模型而发生变化，为了保证KT条件能够持续成立，Cauwenberghs等人通过改造其KT条件的目标函数，使得SVM模型能够跟随新加入的数据动态更新[18]。该算法对于在线系统与流式数据的支持性，这符合社交媒体流数据的要求。

首先记：

于是有：

是为了满足模型能够动态更新而加入的参数，初始值为0，随着新加入模型的向量进行更新。由于需要保证，所以边界向量的参数必须要满足：

其中为非半定矩阵对称矩阵：

可以记：

边界向量需要满足条件消去之后可以变形为：

其中

于是，可以变形为：

记：

由于在KT条件下需要满足：

所以边界向量需要满足：

Online SVM算法将数据以流的形式加入模型，或者将数据从模型中删除，其加入与删除规则具体如下[18]。

每一次向模型中加入新到来的数据时，对于与的更新：

更新的步骤存在有反过程，即数据淘汰过程：

综上总结，每一次向模型中加入新到来的数据时，计算其以及，并且对于重新计算的中间结果与需要考虑如下情况，并且按照如下流程更新：

1. 令

2. ，丢弃新到来的数据

3. 如果 ，按照下面规则使得顺序使得增量最大：

i. ，将新来的数据加入边界向量（支持向量），更新，结束

ii. ，将新来的数据加入错误向量集，结束

iii. *j*保持在内，当左边等号成立从加入正常分类数据集，当右边等号成立时，从加入

iv. , 保持在内，当等号成立时从加入

v. , 保持在内，当等号成立时从加入

#### 改进淘汰策略的在线SVM

仅考虑更新的算法还无法适应本研究的要求，由于社交网络的舆论情感评价会有明显的涟漪效应[31]和淡忘效应[32]。涟漪效应是指舆论事件影响会随着传播不断扩大，针对该事件的情感会越来越强烈。即随着情感数据加入速度越来越快，模型受到新到来数据的影响会越明显。淡忘效应是指舆论事件随着讨论的冷却而迅速被公众遗忘。即数据淘汰随着时间逐渐明显。所以本论文考虑新加入的模型的数据对原模型的影响程度，以及历史数据的淘汰速度。

首先针对于不同时期的数据以及不同影响力的作者的文章对于舆论情感的影响，引入自适应淘汰变量，定义：

为该文章的影响力，同样一篇文章，其在社交网络中的“话语权”越重要，认为该文章对于其评价对象的影响越明显。为了方便模型计算，影响力取为该账号的有效好友数与推文的点赞数转发数之和。

描述为一篇文章在社交媒体中的影响力随时间的变化。文章从发布到被普遍阅读其传播前半段近似指数分布，随着社交圈的扩散达到饱和，最后随着时间衰减逐步被淡忘。特点比较服从Gamma分布[33,34]，故考虑采用Gamma分布进行拟合：

其中：

不同参数的Gamma分布如图5-1所示：

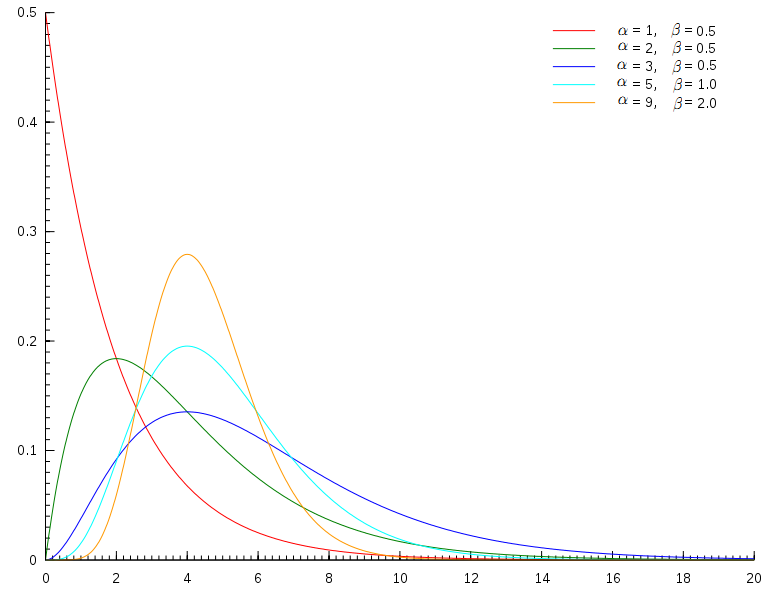


图5-1 不同参数下Gamma过程的图像

参数决定了到达影响力最大值时的时间，参数决定了到达的最大影响力。

在拟合的训练阶段，定义任意时刻的某个话题的影响力数据为：

其中：

为当前话题随时间变化的整体情感评价的真实值。拟合问题本身是一个非线性函数最小二乘拟合问题[35]。即：

设是解的阶近似，考虑在附近的一阶Taylor展开式：

令

记：

原问题（5-9）即可化为：

直接求解目标函数（5-16）梯度为0的点，即目标函数在处的最小值点：

由于不是列满秩矩阵，考虑使用Levenberg–Marquardt算法[35]使用代替. 为LM算法的惩罚因子，于是解得：

记：

所以获取最小化目标函数的参数 的方法：

输入：初始值, 收敛误差，令 0

1. 选取, 计算

2. 计算=

3. 计算与比较，大于则回到1，小于则结束

结束

的选择采用信赖域方法求解[35]：

定义

其中为的第次近似。

*s*

即通过监控步长质量：

初始时选择尽量小，如果步长，那么，迭代点失效，需要缩小信赖半径重新求解。接近1，新迭代点效果较好可以继续，下一次可以增加信赖半径，其余情况保持不变。

在拟合了Gamma函数之后，考虑按照一定的时间间隔，维护每一条在模型中参与计算的文章情感得分自适应淘汰变量。由于淘汰变量的数学定义，使得淘汰变量本身服从正态分布，越靠正态曲线左侧的数据在模型中越不重要，越靠正态曲线右侧的的数据在模型中越重要。根据Gamma函数的特性，随着时间的推进，一条数据在模型中贡献会快速由普通到最大，然后渐渐不重要。

为了得到淘汰变量决定的具体的淘汰策略，首先通过极大似然估计法[36]，计算出正态分布的参数估计：

记：

定义数据最大容量：

其中为当前话题的默认允许的文章数量，有效时间为，选取淘汰阈值为，数据池为.

按照如下规则选择进行更新：

开始

1. 每一次时间窗口，将模型中超过时间的文章数据淘汰出有效推特，重新计算有效推特数量，重新计算一次参数估计，

2. 每一次时间窗口，的数据进行模型淘汰

3. 每一次时间窗口，将新来数据加入数据池，淘汰数据池中超过时间 的已经过期的数据。将数据池中的数据会加入模型更新，没有加入模型的数据重新加入数据池，如果此时模型中的数据大于*V*，则将模型中的按照从小到大的顺序进行末位模型淘汰。

结束

通过调整*k*的大小可以调整参与影响文章的数量，*k*取比较大时，其参与影响模型的总数量会比较多，*k*取比较小时，每次末位淘汰的数据会比较多，会影响使得的向正态曲线右侧移动，会使得加入的数据要求更严格。所以*k*比较小时重新计算的复杂度会降低。

通过调整的大小可以调整不同话题热度的变化速度，的取值一般应该与Gamma函数的保持相对一致。可以取与保持线性的关系。

通过调整*m*的大小可以调整加入模型数据的准入阈值，阈值较大时对于模型的计算量会小，参与计算的数据的情感倾向会普遍比较明显，偏向中立的情感评价加入的概率会变小。阈值较小时，参与计算的数据计算量会变多，计算数据的情感倾向对于整个平台而言更有代表性。

每一次时间窗口在完成更新之后会计算出每一篇文章数据的情感得分，通过计算其社交影响力的加权平均值，可以计算出该话题对应的上市公司最终的情感评分。

#### Passive Aggressive SVM算法的设计

利用动态淘汰策略的在线SVM算法，可以自适应的调整模型的更改速度，而不用重新计算所有数据。但是该算法存在由于参数设置问题导致模型更新过快，旧的数据尚未产生其足够的情感影响即被淘汰的情况。该算法也会出现历史数据积累的效应无法在时间上对模型产生完整影响的情况。在此基础上，本文提出使用一种新的在线被动攻击SVM算法，通过对目标函数进行变形，从而能满足在一定条件下最小化每一条数据对模型产生的影响，且历史数据的影响不被快速覆盖。

将当前时刻现有所有的对于该对象的语言评价的情感得分以及该时间内新到来的的某一条或多条评价进行组合定义，即当前时刻所有的评价矩阵：

为了达到满足一定条件下最小化当前时刻每一条数据对模型产生的影响的效果，即最小化，可以认为找到最小化的目标函数

其中，为模型中允许犯错的松弛变量，为被允许加入模型数据，为被拒绝加入模型或者从模型中淘汰的数据。

对于此类目标函数，可以使用在线被动攻击算法[37] (Online Passive Aggressive Algorithms, Crammer)变形为迭代方式优化，即每一步需找到

其中

目标方程是为了保证能够尽可能的高效分类，同时应当尽可能和保持相近，以便能保留先前步的结果。目标方程的拉格朗日函数为：

其中为函数的拉格朗日乘子，求出*L*对于*Q*, 的偏微分为0的点：

不妨令：

得到：

并且有：

得到

重新代入方程，可以得到：

令不为0，常数替换为, 可以得到：

其中

于是可以得到算法更新的过程如下：

最小化模型更新的Online Passive Aggressive SVM

输入：用户*U*所有已有推文对于对象*O*的评价矩阵*W*，新到来的评价，未加入或者被淘汰的评价 ，参数*C*

输出：*Q*

1. 初始化：为一个半正定矩阵，不妨假设为单位矩阵

2. for t=1, 2, 3, …

计算

计算

计算

计算,其中

结束

###### 时间序列股价预测算法

由于股价在一定程度上服从随机游走模型，其均值来源于市场对于该上市公司的表现预估，方差基于市场的波动情况，在短时间内不考虑非理性因素的情况下，可以认为股价服从随机分布。一般采取时间序列模型[23]对股价进行预测。而时间序列模型一般较多地采用指数平滑法。由于一般可以认为在短期时间内的正常情况下，股价的自然变化没有明显的趋势和倾向，只有在舆论情感评价产生显著变化的情况下才会对股价变化产生影响，故本文采取一阶指数平滑作为基本模型，并引入舆论情感评价因素作为影响股价的变量。

假设上一周期的股价为：,上一周期的预测值为,舆论情感评价造成的影响为,那么第*q+*1周期其一阶指数平滑的预测结果为：

考虑到股价不可能无限下跌也不可能无限增涨。而由于时间窗口的选择往往远小于舆论传播以及发展的时间。所以相对于预测窗口的时间长度，某一对象的舆论情感评价可能处于长期持续上扬或者低迷的状态，因此可能会使得预测结果一直处于单一的上涨或者下跌的状态。市场规律指出，实际的股票市场中，一只股票上涨或者下跌的趋势最后由市场表现决定，一段时间的持久的增长会一定程度减缓增长的速率，所以需要考虑历史增长对于预测结果的影响[7]。

根据股票市场的经验，单一一支发行较久的股票，在短时间内出现超过100%的涨幅十分罕见，可以定义涨跌的相对前个窗口的最高价的比例为归一化变量. 即：

一般认为短期的历史涨跌的幅度对于未来的市场预期的影响符合近似sigmoid的特性[7]，即：

*k*为股价值（10美元到1000美元之间不等）以预测结果之间的线性调和系数，*Q*为情感分析结果。

于是可以得到：

其中，为在时间窗口为q的时刻，针对于对象的情感分析结果。

算法的更新过程如下：

基于情感分析的时间序列股价预测算法

输入：针对于每一个评价对象，每一个时间窗口*q*的情感评价，上一个时间窗口的股价，参数*a,b*，调和系数*k*，影响的时间窗口数

输出：股价的预测结果

1. 初始化初始化为历史30天的平均股价

2. 计算,若（极端情况下认为涨幅和接近100%相同），取。计算当前时刻q对于对象产生的情感评价而造成的股价波动

3. 得到预测值

4. 下一时刻令 ，回到2继续更新股价

结束

参变量*a, b*的组合有多种可以选择，最开始可选取，如果一家公司的股票走势趋于稳健，可考虑增大*a.* 如果一家公司在经济活动上比较激进冒进，风评容易出现比较大的变化，可以考虑增大*b*. 针对不同的公司，参数的选择不尽相同。

## 算法结果

#### 数据集选取

选取的数据集主要包括：

1. 有关电子产品的社交情感评价[39]：数据集共11000条，涉及到来自于twitter的社交网络用户对于iPad，iPhone，Google App，Android Devices等数码相关的产品的情感评价，可能会对Apple，Google以及其他安卓智能设备厂商的股价产生一定影响；

2. 有关航空公司的社交情感评价[40]：数据集共14610条，涉及到来自于 twitter的社交网络用户对于美国各大航空公司的情感评价，对于各大航空公司的股价可能会产生一定影响；

#### 算法比较

为了评估提出的在线情感分析股价预测算法的性能表现，选择Bollen等人使用的GPOMS（）与自组织模糊神经网络[Bolen28]算法效果进行对比比较。

GPOMS将文本转化为"Calm","Alert","Sure","Vital","Kind","Happy"6个可衡量的情感维度，Bollen等人首先通过格兰杰因果关系检验（Granger Causality Test）检测了6个情感维度与道琼斯指数的相关性。但是格兰杰因果关系检验只能猜测其相关性而无法证明其因果性，并且无法验证非线性的相关性，所以选择基于GPOMS的SOFNN模型作为待比较的预测模型。

选择的自组织模糊神经网络是一个五层的混合神经网络，能够在迭代学习过程中自我组织神经元。由于Bollen等人研究发现仅仅"Calm"维度GPOMS评分与股价关系较为密切，所以为了预测时间窗口q时刻的股价，选择过去3天的历史股价与GPOMS的"Calm"维度时间序列数据作为输入，选择\delta=0.04, \sigma=0.01, k\_{rmse}=0.05, k\_d(i)=0.1 (i =1,2,..,r 为输入数据的层数，k\_{rmse}为预定义的训练均方根误差)。

对于更新淘汰策略的在线SVM算法，选取的参数

|  |  |
| --- | --- |
| 参数类型 | 参数值 |
| Gamma | 0.4 |
| Gamma | 0.6 |
| 最大容量比例 k | 1.5 |
| 有效时间 | 7(d) |
| 淘汰速度 | 0.9544 |

其中与以及有效时间选取主要考虑到实际社交网络的中一个话题的涟漪效应以及淡忘效应的速度以及影响，最大容量比例选取主要为了防止数据集过大影响计算速度，淘汰速度取了，认为不满足模型要求的数据处于正态分布的区间之外。

对于在线被动攻击SVM算法，选取松弛因子的系数C=0.023，Q\_1选取为单位矩阵。松弛因子的系数C决定了允许犯错的程度，C越大收敛速度越快，同时允许引入的误差变大。Q\_1的选择主要考虑到计算的简洁。

对时间序列股价预测算法参数选择。不同市场风格公司，例如对于偏向保守的传统航空行业的公司选取参数主要如表6-7所示

表6-7 时间序列的股价预测的参数选择

|  |  |
| --- | --- |
| 参数类型 | 参数值 |
| 调和系数k | 0.43 |
| a | 0.6 |
| b | 1.5 |
| 有效时间 | 7(d) |

以航空公司数据集为例，通过衡量平均绝对百分误差可以得到，提出的两种算法与SOFNN算法的比较如下：

Airlines：1，2，3，4，5，6

MAPE：Online1,Online2,SOFNN

\sigma^2: Online1,Online2,SOFNN

以XX航空公司为例，预测值与走向如图所示

## 结论

## 参考文献