基于深层融合的股票文本主题识别

张加惠 陈致远 赵 峰 安志勇 谢青松
(山东工商学院计算机科学与技术学院 山东 烟台 264005)

摘 要 股票市场在资本市场中占据着重要地位,是经济的晴雨表。专家对股票的评论是投资者进行投资决策的重要依据。因此,如何快速有效地捕获众多专家股评的主题信息,成为股票研究领域的热点。然而目前大多数股票文本主题识别算法,其特征选择方法及分类模式多采用单一的标准。一般而言,单一的标准只能从某个侧面反映文本主题的识别效果,无法全面捕获目标的主体特征。事实上,不同的特征选择标准及分类器模型从不同侧面去理解文本,捕获的特征信息具有较强的互补性。为了提高股票文本主题识别的准确性,文章从信息融合的角度对股票文本进行了多层面融合:1)特征选择层,对多种特征选择方法进行加权融合,使其能够全面表征股票文本的特点;2)决策层,基于SVM-score,对多个分类器进行决策层融合,使其能够提高文本识别的准确性。基于实测数据的实验表明:相比单一模式的文本主题识别方法,文章提出的多层融合算法的识别精度明显更高。

关键词 特征选择,特征融合,SVM-score,文本分类,主题识别
中图法分类号 TP391 **文献标识码** A

Stock Text Theme Recognition Based on Deep Fusion

ZHANG Jia-hui CHEN Zhi-yuan ZHAO Feng AN Zhi-yong XIE Qing-song
(School of Computer Science and Technology, Shandong Technology and Business University, Yantai, Shandong 264005, China)

Abstract The stock market occupies an important position in the capital market and is a barometer of the economy. Experts' comments on stocks are an important basis for investors to make investment decisions. Therefore, how to quickly and effectively capture the subject information of many expert stock reviews has become a hot spot in the field of stock research. However, most stock text topic recognition algorithms currently use a single standard for their feature selection methods and classification models. In general, a single standard can only reflect the recognition of a text topic from one side, and cannot fully capture the subject's main features. In fact, different feature selection criteria and classifier models understand the text from different sides, and the captured feature information has strong complementarity. To this end, in order to improve the accuracy of the theme recognition of stock texts, this paper has a multi-faceted fusion of stock texts from the perspective of information fusion, it includes: 1) Feature selection layer, which performs weighted fusion on multiple feature selection methods to enable it to fully characterize stock text features; 2) The decision-making layer, based on SVM-score, performs decision-making layer fusion on multiple classifiers, which can improve the accuracy of text recognition. Experiments based on measured data show that the recognition accuracy of the multi-layer fusion algorithm proposed in this paper is significantly improved compared with the single-mode text topic recognition method.

Keywords Feature selection, Feature fusion, SVM-score, Text categorization, Subject recognition

1 引言

股票市场在资本市场中占有极其重要的地位,不仅给投资者带来利益,也极大地促进了国家经济的发展,因此,股票市场一直受到投资者和专家的关注和重视。股票市场一旦发生变化,不仅会影响到普通大众的日常生活,还会对经济的发展产生重大的影响,因而被人们称为经济的"晴雨表"和"报警器"[1]。对股票的准确预测,不仅有利于投资者控制风险,获取利益,还可以帮助政府和经济部门及时掌握、引导和调控股市的健康发展。专家对股票的评论是众多投资者进行投资决策的重要依据。如何准确、有效地从众多的专家股评中捕获文本的主题信息,提高预测结果,一直是金融界和学术界探索的问题,也是投资者们关注的热点。

针对股票文本,目前已有众多学者从不同视角,采用不同方法对股票涨跌趋势进行预测。例如,梁雪玲[2]运用 NSGA-II 的特征选择方法和 MLP 神经网络的算法来对股票进行交易决策研究;卜乐[3]运用 T 检验的特征选择方法和回归模型来研究上市公司股票股利与长期股票价格的相关性,进一步为上市公司的预测提供依据;汤浩[4]运用卡方检验的特征选择方法、灰色-马尔科夫模型和 B-J 时间序列模型对股票进行

本文受国家自然科学基金(61773244),烟台市重点研发计划项目(2017ZH065,2019XDHZ081),赛尔网络下一代互联网技术创新项目(NGII20170626),山东工商学院研究生科技创新基金项目(3110318)资助。

张加惠(1995—),女,硕士生,主要研究方向为金融数据分析与挖掘;**谢青松**(1965—),男,教授,硕士生导师,CCF 会员,主要研究方向为金融数据分析、智能算法,E-mail:xieqingsong@sdtbu.edu.cn(通信作者)。

价格预测；Kim[5] 探讨了支持向量机（Support Vector Machine，SVM）在股票预测应用中的可行性，并提供了一个有前途的股票市场预测模型；厦门大学方匡南教授[6] 利用 R 型聚类和 C5.0 决策树模型预测股票趋势；李妍[7] 运用相关性分析的方法对数据进行特征选择，分别利用 BP-NN，SVM，ELM 对股票进行预测；Han 等[8] 采用回声 SVM 算法预测股票涨跌情况。

上述研究中常用的特征选择[9] 方法包括 Fisher score[10]、卡方检验[11]（Chi-square test，CHI）、T 检验等。其中，Fisher score 是基于距离度量的方法，基本思想是当某个特征的类间离散度与类内离散度比值最大时，Fisher 赋予该特征最高的 Fisher 分值；卡方检验是以分布为基础的一种常用假设检验方法，其基本思想是观察并检验统计样本的实际观测值与理论推断值之间的偏离程度，该值决定了卡方值的大小；T 检验，亦称 student t 检验（Student's t test），是最常用的一种假设检验类型，主要验证总体均值间是否存在显著性差异，其基本思想是用分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著。而分类方法中最常用的分类器模型是支持向量机。支持向量机[12] 是 Vapnik 和 Cortes 于 1995 年正式提出的，其基本思想就是找到一个对训练样本局部扰动的“容忍性”最好的划分超平面，使分类结果最鲁棒，对未见示例的淡化能力最强，同时提高对测试样本的预测准确率；核心在于核函数的构造，能够将原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分，并且计算量不会增加。

已有文本主题识别研究算法多是利用单一的特征选择方法来提取文本特征，并且使用一个分类器模型来进行文本识别。这种单一的主题识别模式只能从某个侧面描述文本的主题特征，无法全面捕获目标的主体特征。不同的特征选择标准及分类器模型捕获的特征信息具有较强的互补性。因此，采用特征融合技术整合目标的多视角特征，成为客观、全面、准确描述目标特性的主要手段。

受融合原理的启发，本文首先从不同视角抽取文本的特征并进行加权融合，然后基于 SVM 的 score 得分进行决策层融合，构建了一个增强分类器用于最终的文本主题判别。图 1 给出了本文多层次融合的流程，具体为：1）文本预处理，通过爬取东方财富网上的专家股评，对文本进行分词、去停用词的处理，并进行文本表示，以降低文本的复杂性；2）特征选择，结合特征选择原理和方法，运用加权融合的思想，抽取判别性较强的融合特征，能够全面、准确地描述文本特点；3）分类决策，通过 SVM 的 score 得分对融合后的特征进行加权融合，构建 SVM 增强分类器，实现对股票文本的主题识别。

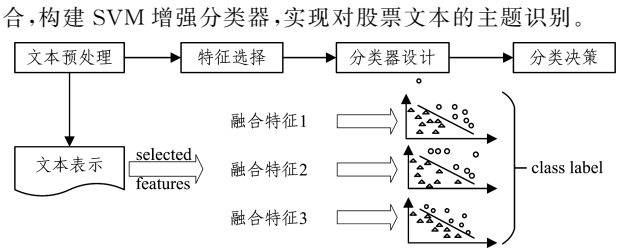

图 1　层次融合流程图

本文第 2 节介绍文本预处理的基本流程和方法；第 3 节对特征选择层融合进行相关概述；第 4 节概述分类决策的相关技术，并简单介绍本文的具体融合算法；第 5 节给出实验结果及其分析，针对本文提出的融合算法进行实测数据的实验，并且将其与单一模式的文本主题识别方法进行比较；最后总结全文，并针对未解决的问题提出了下一步的规划和展望。

2　数据获取及处理

数据来源于东方财富网[1)]。本文选取了 2018 年 10 月 4 日至 2018 年 10 月 17 日期间所有专家对“大盘分析”的股评文章，共计 100 篇。然后对其进行文本预处理，流程如图 2 所示，具体包括：1）对文本进行分词[13]，本文选取的是 Python 中的 Jieba 分词工具；2）去停用词，本文选用的是哈工大停用词表[14]；3）文本表示，目前应用较多且效果较好的方法主要是向量空间模型[15]、布尔逻辑模型[16]、概率模型[17] 等。本文选用的是布尔逻辑模型，其表示形式如下：

$$X=\begin{bmatrix} x_{11}^{k} & \cdots & x_{1n}^{k} \\ \vdots & & \vdots \\ x_{m1}^{k} & \cdots & x_{mn}^{k} \end{bmatrix} \tag{1}$$

其中，$i=1,2,\cdots,m$ 表示第 i 个文本特征，$j=1,2,\cdots,n$ 表示第 j 篇文本，$k=1,2,\cdots,c$ 表示文本属于的类别。

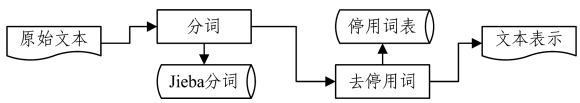

图 2　文本预处理流程图

3　特征选择层融合

本文选用的特征选择方法是 Fisher score、卡方检验、T 检验。

3.1　特征选择方法

Fisher score 是基于距离度量的方法，基本思想是根据 Fisher 判别准则计算特征的比值，并将其作为该特征的 Fisher 分值，值越大，表明特征的分类能力越强[18]。其数学描述为：对于分类问题，给定训练样本 (x_{rj})，$x\in R^{d}$，d 为原始特征空间维数，样本个数记为 $l_k\ (k=1,2,\cdots,c)$，则第 r 个特征的 Fisher 分值为：

$$F(r)=S_{br}/S_{wr} \tag{2}$$

其中，S_{br} 为第 r 个特征的不同类之间的离散度，描述样本间的距离；S_{wr} 为第 r 个特征的同一类之间的离散度，描述同类样本间的距离。其计算公式为：

$$S_{br}=\sum_{k=1}^{c}(\overline{m_{kr}}-\overline{m_r})^2 \tag{3}$$

$$S_{wr}=\sum_{k=1}^{c}\frac{1}{l_k}\sum_{x_r\in X_k}(x_r-\overline{m_{kr}})^2=\sum_{k=1}^{c}\delta_{kr}^2 \tag{4}$$

$$\overline{m_{kr}}=\frac{1}{l_k}\sum_{x_{kr}\in X_k}x_{kr},k=1,2,\cdots,c \tag{5}$$

卡方检验是以 χ^2 分布为基础的一种常用假设检验方法，其基本思想是观察并检验统计样本的实际观测值与理论推断值之间的偏离程度，该值决定卡方值的大小。其计算公式为：

$$\chi^2(r)=\sum\frac{(O(r)-E(r))^2}{E(r)} \tag{6}$$

$$E(r)=\frac{N_1*M_1}{n} \tag{7}$$

其中，$O(r)$ 指特征 r 的实际次数或观测次数，$E(r)$ 指特征 r 的期望次数或理论次数，N_1 表示列总数，M_1 表示行总数，n 为

1) http://quote.eastmoney.com/zs000001.html

样本总量。

T 检验是最常用的一种假设检验类型，主要验证总体均值间是否存在显著性不同，其基本思想是用分布理论来推论差异产生的概率，从而检验两个总体的均值差异是否显著。

设 n_1 个文本来自正态总体 $N(\mu_x,\sigma_x^2)$，n_2 个文本来自正态总体 $N(\mu_y,\sigma_y^2)$，两组样本的均值为 $\overline{m_{n_1}}$ 和 $\overline{m_{n_2}}$，方差分别为 δ_{n1}^2 和 δ_{n2}^2。第 r 个特征的计算公式为：

$$T(r)=\frac{(\overline{m_{n_1}}-\overline{m_{n_2}})-(\mu_x-\mu_y)}{\delta_r\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{8}$$

$$\delta_r^2=\frac{(m_1-1)\delta_{n_1}^2+(m_2-1)\delta_{n_2}^2}{n_1+n_2-2} \tag{9}$$

将 T 检验用于多分类时，其计算公式为：

$$T(r)=\frac{\sum_{k_1=1}^{c}\sum_{k_2=1}^{c}T_{k_1k_2}(r)}{C_k^2} \tag{10}$$

其中，k_1 和 k_2 代表文章所属的类别。

3.2 特征选择融合

上述单一的特征选择方法只能从某个侧面描述文本的主题特征，无法全面捕获目标的主体特征。而基于不同的特征选择标准所捕获的特征信息具有较强的互补性，因此采用特征融合技术整合目标的多视角特征，是客观、全面、准确描述目标特性的主要手段。本文在上述几种特征选择方法的基础之上，提出了一种加权融合方法，其核心思想是：从不同视角，按照一定的原则给每个特征值定制加权因子 β，最后加权综合所有得到全局特征值，融合后的加权因子满足 $\sum_{i=1}^{p}\beta_i=1$，即 $\beta_1+\beta_2+\cdots+\beta_p=1$。不同加权因子的选择，对文本分类的识别效果不同，选择合适的加权因子能够对文本主题识别达到极佳的效果。这里，一定的原则是指在[0,1]的范围内以 0 为开始，0.1 为步长，1 为结束，给特征选择方法赋予加权因子值进行特征融合。经过加权融合后，将具有最佳分类效果的参数用于构建最优值和测试数据。

特征选择加权示意图如图 3 所示，设 Fisher score 的加权因子为 β_1，卡方检验的加权因子为 β_2，T 检验的加权因子为 β_3，加权融合后的第 r 个特征值为：

$$R(r)=\beta_1F(r)+\beta_2\chi^2(r)+\beta_3T(r) \tag{11}$$

其中，$\beta_1+\beta_2+\beta_3=1$，$F(r)$为特征 r 的 Fisher 分值 0，$\chi^2(r)$为特征 r 的卡方值，$T(r)$为特征 r 的 T 检验值。为了获得一组更有意义和有辨别力的特征，我们建议使用加权融合系数的修改版来量化每个文本特征的细微差别。

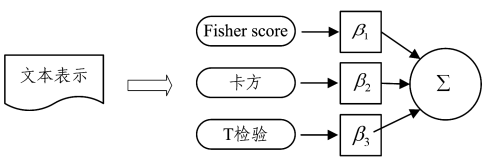

图 3 特征选择加权示意图

4 分类决策

SVM 寻求最大边缘超平面来将一个类的样本与另一个类分开。训练数据的经验风险和模型的复杂性可以是超参数，从而确保对看不见的数据具有良好的泛化能力。

对特征选择加权融合完后的特征，进一步基于 SVM 的 score 得分进行决策层融合，构建一个增强分类器用于最终的文本主题判别。score 得分反映了点到边缘的距离，值越大，表示 score 属于该类的可能性越大。最终的分类结果是通过融合所有 SVM 的 score 得分获得的。决策层融合的模型如图 4 所示。

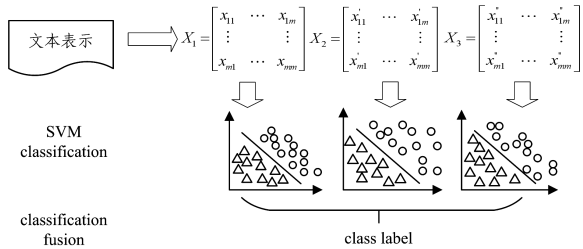

图 4 决策层融合示意图

我们训练 $h(h=1,2,\cdots,z)$个 SVM 分类器，每个分类器训练一个特定的特征集 $\{\overline{\gamma_i}\}_{i=1}^{h}$，其中 h 表示用于计算不同级别的功能连接的级别数。然后，来自所有 SVM 模型的决策分数线性融合(通过针对每个 SVM 调整的加权参数，加权因子 α 从 0.1 到 0.9 选择，步长为 0.1)以产生目标主题的最终标签，记为 $S\_c_{jh}^k$，其表示第 j 个文本在第 h 个分类器中属于第 k 类的得分。最后，来自所有 SVM 模型的 score 得分线性加权融合，即满足：

$$S\_c_j^k=\sum_{i=1}^{n}\beta_i\sum_{h=1}^{z}\sum_{k=1}^{c}S\_c_{jh}^k \tag{12}$$

通过 $S\_c_j^k$ 的大小来判断第 j 个文本所属的类别，以产生目标对象的最终标签。

本文的具体融合流程如下：1)特征选择层，根据式(2)、式(6)、式(10)计算每个特征的 $F(r)$，$\chi^2(r)$，$T(r)$，利用加权融合方法提取特征(见 3.2 节特征选择融合)；2)选取适当数目的分类器，对 1)中的特征进行分类器设计，并分别计算各自的 score 得分；3)将 score 得分作为决策层的输入，进行决策层的加权融合，获取最终的分类预测标签。

5 实验结果及分析

5.1 数据描述

本文相应的实验文本数据来源于东方财富网[1)]。网站上有大盘分析、证券要闻、名博论市、公司快讯、上证指数吧、研究报告等信息。本文采用爬虫技术来爬取“大盘分析”的相关文本报道。专家对股票的分析文本中，一般包括专家对股票的评价、分析、走向看法、政府政策、建议等。本文随机选取 2018 年 10 月 4 日至 2018 年 10 月 17 日期间的 100 篇文本进行人工手动标签以表示专家对股票未来趋势的预测，趋势一共分为 3 种：看涨(1)、观望(0)、看跌(−1)。其中，代表看涨的观点有 39 篇，代表观望的观点有 30 篇，代表看跌的观点有 31 篇。经过文本预处理后的数据特征，共有 6315 维。在实验中，采用 libsvm[19] 作为训练和测试工具，运用 C-SVR 模型，SVM 中采用线性核函数。

本文融合方法的性能取决于一些超参数，例如在特征选择融合步骤的 β(见式(11))，在 SVM 模型中参数 γ 的取值以及在决策融合步骤中的 α(见式(12))，重要的是微调这些参数。因此，我们对训练数据使用 5 折交叉验证[20]，在以下范

1) http://quote.eastmoney.com/zs000001.html.

围内自动识别这些超参数的最佳值：$\beta\in[0,0.1,\cdots,1]$，$\gamma\in[2^{-5},2^{-4},\cdots,2^{5}]$和 $\alpha\in0,0.1,\cdots,1]$。具体而言，我们进一步将训练集分成训练子集和验证子集，并进一步执行交叉验证。也就是说，对于超参数的每个值组合，来自训练集的验证子集用于测试，剩余的训练子集用于训练。该过程重复 10 次，并在超参数值的特定组合下产生分类准确度。然后，选择在验证数据上具有最佳分类精度的超参数值，并基于所有训练样本构建最优模型，将具有优化参数的构建模型应用于测试数据。此过程中，我们采用两种不同的统计学指标，即分类识别准确率(ACC)和识别率方差(VAR)。准确率越高，反映文本主题识别越优良；方差越小，反映主题识别效果越稳定。

5.2　不同特征选择标准特性的对比分析

根据本文的特征选择方法，按照式(2)、式(6)、式(10)计算各个特征的 Fisher 值、卡方值和 T 检验值，进行大小排序后，将特征加入到特征集中，然后对其进行文本识别，其特征选择数据见图 5，识别结果见表 1；然后按照 3.2 节所讲述的融合方法进行加权融合，采用 5 折交叉验证来实现文本主题的识别，重复 10 次，并将其与单一的特征选择标准进行对比，其最优特征选择数据见图 6，其结果见表 1，其中“+”代表将两种方法加权融合。

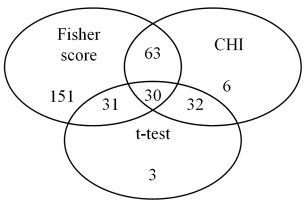

图 5　单一特征选择标准数据图

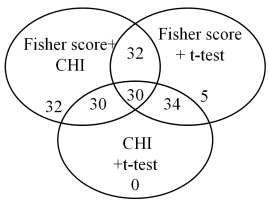

图 6　特征选择融合数据图

通过图 5 可以看出，将文本预处理后的数据按照上述过程进行单一的特征选择主题识别时，从不同的角度提取的特征是不一样的，选择的特征数目也是不一样的。Fisher score 的最优特征数据为 214，卡方检验的最优特征数据为 69，T 检验的最优特征数据为 35，其交集部分说明了不同的特征选择标准选择的数据具有一定的相同项，非重叠的部分说明了数据还存在较大的差异性。图 6 为特征选择加权融合后的最优数据，Fisher score 和卡方检验融合后的最优特征数据为 64，Fisher score 和 T 检验融合后的最优特征数据为 37，卡方检验和 T 检验融合后的最优数据为 34。相比单一的特征选择标准，交集增多，非交集部分减少，说明选取的相同特征增多，但数据还存在一定的差异性，识别效果还有待增强，因此需要在此基础之上进行分类决策层的融合来进一步减少差异。

表 1　特征选择加权融合实验结果

特征选择方法	数据	SVM 参数	*ACC*	*VAR*
Fisher score	214	1×2^{-4}	74.00	1.30
卡方	69	1×2^{-4}	73.00	1.20
t-test	35	1×2^{-3}	74.50	0.80
Fisher score＋卡方	64	1×2^{-4}	76.00	0.70
Fisher score＋t-test	37	1×2^{-4}	77.00	0.80
卡方＋t-test	34	1×2^{-4}	77.00	0.50

表 1 对单一的特征选择方法和融合后的特征选择方法进行了文本主题识别对比，产生了最优的结果和参数。从表中可以看出单一的特征选择标准的识别准确度相近，但选取的特征数据存在较大的差异性，因此进行特征选择融合来减少数据差异性，提高股票文本主题的识别率，具有必要性。而进行特征选择融合后的文本主题识别，准确率明显提高，识别效果相对稳定。单一特征选择方法和特征选择融合后的最优结果对比表明：相比单一模式的特征选择方法，本文提出的特征选择融合算法的识别精度明显提高，通过方差的大小可以看出融合后的特征的稳定性明显有了提高。

5.3　参数对识别结果的影响

图 7 所示的折线图为特征数目对识别准确率的影响。从图中可以看出，选择不同的特征数目对识别准确率有不同的影响。选择合适的特征数目是进行文本主题识别、提高识别准确率的关键。

图 8 所示的折线图为特征选择融合权重对识别准确率的影响。从图中可以看出，不同的权重对识别准确率的影响不同，选择合适的权重对文本主题识别有不同的分类准确性。其中，Fisher score 和卡方检验融合的最佳权重为(Fisher score：0.2，卡方：0.8)；Fisher score 和 T 检验融合的最佳权重为(Fisher score：0.1；T 检验：0.9)；卡方检验和 T 检验融合的最佳权重为(卡方检验：0.1，T 检验：0.9)。

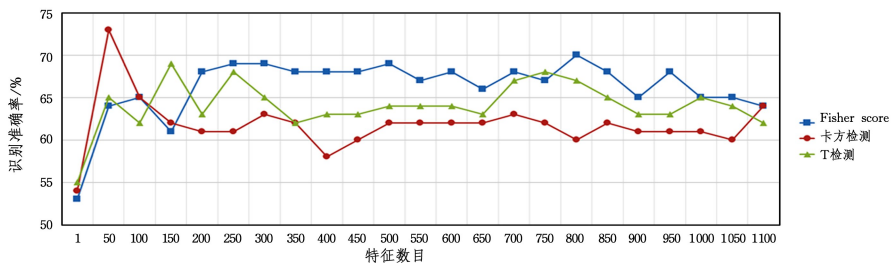

图 7　特征数目对识别准确率的影响

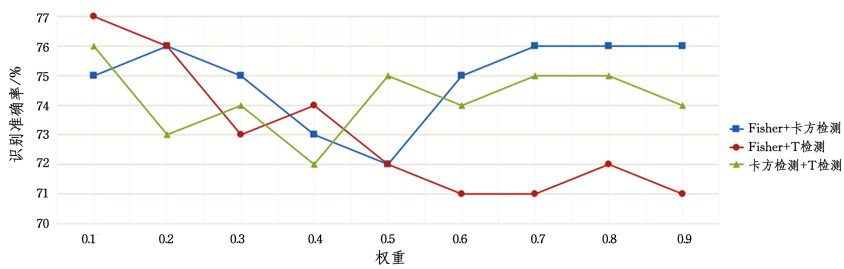

图 8　权重对识别准确率的影响

图 9 为分类决策融合下的平均分类精度，平均分类精度(ACC)由相应的直方图长度和颜色编码。精度越高，长度越长，颜色越暖。其中，X 轴 α_1 表示的是 Fisher score 和卡方检验融合后的 score 得分的权重；Y 轴 α_2 表示的是 Fisher score 和 T 检验融合后的 score 得分的权重；相应地，$\alpha_3=1-\alpha_1-\alpha_2$，$Z$ 轴 ACC 表示的是识别准确率。从图中可以看出，经过融合后的最佳平均分类精度为 80，最小分类精度为 74。正是由于平均分类精度对值比较敏感，所以一个合适的参数对分类精度非常重要。通过实验比较发现，$\alpha_1=0$，$\alpha_2=0.9$，$\alpha_3=0.1$ 是加权融合的最佳参数。

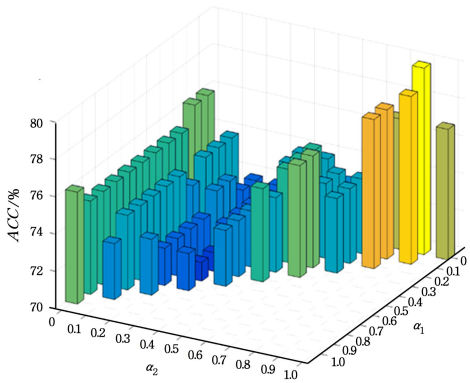

图 9 不同参数融合下的识别准确率

5.4 分类决策融合的结果及分析

根据本文的特征选择融合算法，在计算出每个特征的加权值并进行大小排序后，将大的特征值加入到特征集中，进一步基于 SVM 的 score 得分进行决策层的加权融合，并查看最终的识别率，结果见表 2，其中“+”代表将其进行融合。

表 2 分类决策融合实验结果

融合	权重	*ACC*	*VAR*
FK+FT	(0.8,0.2)	79.00	0.40
FK+KT	(0.1,0.9)	79.00	0.25
FT+KT	(0.3,0.7)	78.00	0.30
FK+FT+KT	(0,0.9,0.1)	80.00	0.15

通过表 2 的实验数据可以看出，在上述基于特征选择融合的基础上，进一步基于 SVM 的 score 得分进行分类决策融合，其文本主题识别率明显提高。综合所有实测数据的实验分析，我们得出以下结论：1)特征选择标准之间确实存在一定的差异性；2)将不同的特征选择标准进行加权融合，在一定程度上能消除单一模式存在的差异性；3)特征选择加权融合能有效提高文本主题识别的准确率；4)基于 SVM-score 得分的文本主题识别算法的识别准确率明显比特征选择融合的准确率高；5)相比单一模式的文本主题识别方法，本文提出的多层融合算法的识别精度有了明显提高。

结束语 文本是股票预测的重要依据，是引导投资者做出更好的投资、减少投资风险的保证。针对文本冗余和分类准确率低的问题，本文提出了一种基于特征选择和分类决策双层融合的分类算法。该算法通过计算每个特征的各个特征选择方法值，将其进行加权融合，然后利用支持向量机的决策层进行进一步融合来实现文本主题识别。实验结果表明：本文的融合算法具有较好的分类结果和较高的识别率。在未来的工作中，将探究特征提取方法对文本的主题识别和分类效果，以获得更好的特征，进一步提高对股票文本主题识别的准确度。

参 考 文 献

[1] 张晨希. 数据挖掘技术在股票预测中的应用[D]. 合肥：安徽大学，2006.

[2] 梁雪玲. LG-trader：基于局部泛化误差和特征选择的股票交易决策支持[D]. 广州：华南理工大学，2014.

[3] 卜乐. 我国上市公司股票股利与长期股票价格相关性研究[D]. 上海：东华大学，2014.

[4] 汤浩. 股票收益分布函数分析及价格预测[D]. 武汉：武汉科技大学，2004.

[5] KIM K J. Financial time series forecasting using support vector machines [J]. Neurocomputing，2003，55(1)：307-319.

[6] 方匡南，纪宏，路逊. 股票技术指标相似性与有效性研究[J]. 统计与信息论坛，2009，24(9)：26-30.

[7] 李妍. 基于集成学习的股票买卖点预测研究[D]. 西安：西北大学，2018.

[8] HAN M，XI J H，XU S G. Prediction of chaotic time series based on the recurrent predictor neural network[J]. IEEE Transactions on Signal Processing，2004，52(12)：3409-3416.

[9] GUYON I，ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research，2003(3)：1157-1182.

[10] 张润莲，张昭，彭小金，等. 基于 Fisher 分和支持向量机的特征选择算法[J]. 计算机工程与设计，2014，35(12)：4145-4190.

[11] 宋哲理，王超，王振飞. 基于 MapReduce 的多级特征选择机制[J]. 计算机科学，2018，45(S2)：478-483，489.

[12] MAO X，ZHAO G，SUN R. Naive Bayesian algorithm classification model with local attribute weighted based on KNN [C]// Proc of IEEE Information Technology，Networking，Electronic and Automation Control Conference. IEEE，2017：904-908.

[13] 汪东升，黄传河，黄晓鹏，等. 电信大数据文本挖掘算法及应用[J]. 计算机科学，2017(12)：238-244.

[14] 数据堂. 停用词集合[DB/OL]. http://www.datatang.com/data/19300/. Data Hall. Stop word collection[DB/OL]. http://www.datatang.com/data/19300/ .

[15] 王纵虎，刘速. 一种成对约束限制的半监督文本聚类算法[J]. 计算机科学，2016，43(12)：190-195.

[16] 李荣陆. 文本分类及其相关技术研究[D]. 上海：复旦大学，2005.

[17] 刘付勇，高贤强，张著. 基于改进贝叶斯概率模型的推荐算法[J]. 计算机科学，2017，44(5)：285-289.

[18] MESLEH A W. Chi square feature extraction based SVMs Arabic Language Text Categorization system[J]. Journal of Computer Science，2007，3(6)：430-435.

[19] CHANG Ch C，LIN C -J. LIBSVM：A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology，2011，2(27)：1-27.

[20] NASON G P . Wavelet Shrinkage Using Cross-Validation[J]. Journal of the Royal Statistical Society：Series B (Methodological)，1996，58(2)：463-479.