文章编号:1674 - 2869(2019)03 - 0296 - 07

基于AP-SVM组合模型的股票价格预测

胡 迪[1,2],黄 巍*[1,2]
1. 武汉工程大学计算机科学与工程学院,湖北 武汉 430205;
2. 智能机器人湖北省重点实验室(武汉工程大学),湖北 武汉 430205

摘 要:为了减小单支股票训练数据中的噪声对分类器性能的影响,提出了一个新的基于簇的股票价格涨跌预测方法(AP-SVM)。AP-SVM首先使用近邻传播(AP)算法挑选出与待预测股票价格变化相似度较高的其他股票,然后将待预测股票和与其价格变化相似的其他股票一起作为输入数据,训练一个支撑向量机(SVM)实现对待预测股票价格涨跌的预测。实验结果表明,当训练数据中存在噪声时,AP-SVM在预测准确率方面优于传统的SVM方法。

关键词:股票价格预测;近邻传播;支持向量机;AP-SVM模型
中图分类号:TP391 **文献标识码**:A **doi**:10. 3969/j. issn. 1674-2869. 2019. 03. 017

Stock Price Trend Prediction Based On AP-SVM Combined Model

HU Di[1,2], *HUANG Wei*[*1,2]
1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China
2. Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology), Wuhan 430205, China

Abstract: To overcome the performance degradation of the classifier due to the noise in individual stock price data, this paper proposes a novel cluster-based stock price trend prediction method (AP-SVM). We first used the affinity propagation algorithm to select all the stocks with high similarity in the aspect of the price tendency to the predicted stock, and then used these stocks as the training data to fit a support vector machine (SVM). The experimental results show that AP-SVM surpasses the traditional SVM method in predicting accuracy in the noisy training data.

Keywords: stock price trend prediction; affinity propagation; support vector machine; AP-SVM model

股票涨跌趋势预测是利用历史的股票数据信息对未来的股票走势方向做出预测。股票价格受政治,经济和人为操作等许多因素影响,股票的涨跌是这些因素的综合表现。对于投资者来说要考虑这些因素,需要花费大量的人力和物力来收集这些信息,在股票数据信息爆炸式增长的时代,这种做法是不可取的。因此在股票数据上直接进行预测分析,既能避免繁杂的数据获取过程,又能避免大量的冗余信息的干扰。目前,股票预测的方法有很多种,大致可以分为三类:人为经验法,统计学方法和人工智能方法。随着人工智能技术在计算机视觉和自然语言处理领域的表现越来越出色,在时序预测方面相比于人为经验法和统计学方法有更大的优势。

股票涨跌预测是金融领域经久不衰的研究课题,由于人为经验法和统计学方法等传统方法在股票涨跌预测中的效果并不理想,国内外研究者开始探索新的股票预测方法,为了能够准确预测股票涨跌的变化趋势,学者们进行了大量的分析和验证,探索出了越来越多的预测方法和模型。

彭丽芳等[1]提出一种基于时间序列的SVM股票预测方法,建立股票收盘价回归预测模型,克服

收稿日期:2019-03-02
作者简介:胡 迪,硕士研究生。E-mail:944675054@qq.com
***通信作者**:黄 巍,博士,副教授。E-mail:wei.huang@foxmail.com
引文格式:胡迪,黄巍. 基于SVM的聚类股票涨跌预测[J]. 武汉工程大学学报,2019,41(3):296-302.

了传统时间序列预测模型仅局限于线性系统的情况，提高了预测准确率。金桃等[2]提出采用基于支持向量机(support vector machine，SVM)的多变量股市时间序列预测算法来提高预测准确率，实验证明相较于单变量的SVM回归预测有更好的泛化能力。程昌品等[3]提出二进制正交小波变换和自回归平均移动-支持向量机(autoregressive integrate moving average support vector machine，ARIMA-SVM)方法，使用小波分解[4]算法对数据进行分解，分离出非平稳时间序列中的低频和高频信息，然后对高频信息构建自回归平均移动模型(autoregressive integrate moving average，ARIMA)预测，对低频信息用SVM模型进行拟合，最后将各模型的预测结果进行叠加。实验表明，小波分解ARIMA-SVM的组合模型较单一的预测模型效果更高。黄同愿等[5]通过选择最优的径向基核函数，再利用网格寻参[6]，遗传算法[7]和粒子群算法[8]对最佳核函数参数进行对比寻优，构建的SVM模型能够准确的预测股票反转点。张贵生等[9]提出近邻互信息特征选择的支持向量机-广义自回归条件异方差模型(svm support machine-generalized autoregressive conditional heteroskedasticity，SVM-GARCH)，通过近邻互信息的方式融合了与目标指数数据关系密切的周边证券市场的相关变化信息，仿真结果表明在时序数据除噪，趋势判别以及预测的精确度等方面均优于传统的自回归平均移动-广义自回归条件异方差模型(autoregressive integrate moving average-generalized autoregressive conditional heteroskedasti，ARIMA-GARCH)。李辉等[10]提出一种两层特征选取及预测的方法，即特征子集区分度衡量准则-二进制粒子群-支持向量机(discernibility of feature subsets-binary particle swarm optimization-support vector machine，DFS-BPSO-SVM)，第一层特征选取高效剔除部分非预测相关特征，缩减了特征规模，第二层选择出最优特征组合，提升了预测准确率。张伟等[11]提出将SVM和遗传和支持向量机组合算法(genetic algorithm-support vector machine，GA-SVM)，预测未来股票市场的走势，实验证明GA-SVM优于其他方法。胡蓉[12]提出多输出的SVM回归模型，预测股票的最高价和最低价，与单输出相比有更好的整体预测精度和抗噪性能。

曾岫等[13]针对K线图是一个分形图，把其分维数作为聚类参数对股票进行聚类实证研究，研究结果表明，同一类的股票有着极强的相似走势。柯冰等[14]利用9项财务指标，对19家上市公司进行聚类分析，能把上市公司分为4个不同的类，与公司的实际情况相符。由此可见，聚类能将相同走势的股票归类在一起。吴薇等[15]利用反向传播神经网络(back propagation，BP)网络[16]有较好的分类能力，对沪市综合指数涨跌进行预测，实验结果表明BP网络对中国股票市场的预测是可行的和有效的。

过去的研究一般考虑对单支或少数股票进行实证分析，忽略了股票与股票之间相关性，这种相关性的股票之间往往会有相同的变化趋势。本文提出了基于SVM的聚类股票预测算法近邻传播聚类和支持向量机组合算法(affinity propagation-support vector machine，AP-SVM)，首先用近邻传播算法(affinity propagation，AP)对股票进行聚类，然后将簇内股票按时间索引构建成矩阵簇，最后利用SVM，BP，AP-SVM和近邻传播和反向传播组合算法(affinity propagation-back propagation，AP-BP)对太平洋证券进行涨跌预测，对比分析4种算法在不同时间间隔上的预测表现，同时进一步讨论簇内股票数目对于预测结果的影响。

1　AP-SVM模型

A股股票常用特征有7个，即日期、最高价、最低价、开盘价、收盘价、成交量和成交额。首先对除时间以外的其余6个特征做相关性分析，计算各个特征之间的相关系数，选择相关性较弱的特征作为模型的输入，由于这些特征之间的差异性较大，需要先对其进行最大最小归一化处理，然后对收盘价做二范数归一化[17]处理，将处理好的数据进行聚类分析，最后利用不同算法进行训练，预测聚簇股票的涨跌。

1.1　特征选择

由于股票数据中的特征存在相关性，会造成数据冗余，对预测结果产生影响，需要对股票特征之间进行相关性分析，挑选出能够有效代表完整股票数据的特征子集。使用相关系数来描述各个特征之间的相关性，相关系数的取值在[-1，1]之间，-1表示完全负相关，1表示完全正相关，0表示不相关。相关系数n的定义如下：

$$n=\frac{E[(\boldsymbol{v}_i-\mu_i)(\boldsymbol{v}_j-\mu_j)]}{\sqrt{D(\boldsymbol{v}_i)D(\boldsymbol{v}_j)}} \tag{1}$$

其中μ_i和μ_j分别是特征$\boldsymbol{v}_i$和特征$\boldsymbol{v}_j$的均值，E和D分别表示计算期望和方差。

中信证券(股票代码为600030)各特征之间的

相关系数如图1所示。由图1可知，开盘价、收盘价、最高价和最低价之间的相关系数接近1，表示这些特征之间相关性很强，成交量和成交额之间也是相关性很强，而开盘价、收盘价、最高价、最低价与成交量、成交额之间相关性较小。因此只需要2个输入特征就够了，由于涨跌预测需要用到收盘价的差值来计算，根据图1的第二行，收盘价与成交量的相关系数最小，所以选取成交量当做是股票的第二个输入特征。

图1 中信证券特征之间的相关系数

Fig. 1 Correlation coefficient diagram between CITIC securities features

由于股票的收盘价和成交量在数量级上相差较大，对2个特征进行最大最小归一化处理（规范化到[0,1]区间），以减少训练过程中计算的复杂度和预测准确率。归一化的公式如下：

$$x'_i=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{2}$$

其中，$x_{\max}$ 和 $x_{\min}$ 分别为样本中特征的最大值和最小值，x'_i 为归一化后的数据。

1.2 聚类相似度度量

相同走势的股票之间往往有较高的相似性，可以利用股票价格间的相似性提高股票价格预测的准确性。

聚类操作可以挑选出与待预测股票价格走势相近的股票。首先对每支股票的收盘价进行二范数归一化，假设某支股票包含 N 天的收盘价格，则这支股票的收盘价格可记为 $\boldsymbol{x}=(x_1,x_2,\cdots,x_N)$，其二范数归一化定义如下：

$$x_i^*=\frac{x_i}{\sqrt{\sum_{i=1}^{N}x_i^2}} \tag{3}$$

式(3)中，x_i 表示这支股票第 i 天的收盘价，x_i^* 表示这支股票第 i 天的归一化收盘价。二范数归一化如图2所示。

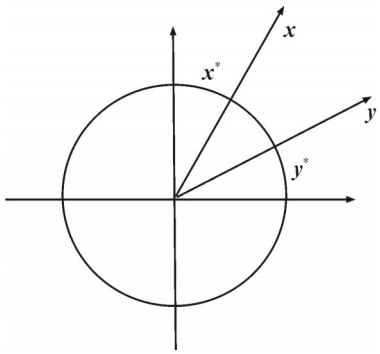

图2 收盘价的二范数归一化

Fig. 2 Normalization of two norms of closing price

在图2中，$\boldsymbol{x}$，$\boldsymbol{y}$ 表示2支不同股票在同一个交易日的原始收盘价，$\boldsymbol{x}^*$，$\boldsymbol{y}^*$ 表示 $\boldsymbol{x}$，$\boldsymbol{y}$ 归一化后股票的收盘价。从图2中可以看出，二范数归一化后，所有股票的收盘价格向量落到一个球面上，因而可以使用2支股票收盘价向量 x 和 y 的夹角余弦刻画它们的相似程度。

由向量内积定义可知，2支股票价格向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的夹角余弦定义如下：

$$\cos\theta=\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2} \tag{4}$$

对于二范数归一化后的股票价格向量 $\boldsymbol{x}^*$ 和 $\boldsymbol{y}^*$，有 $\|\boldsymbol{x}^*\|_2=\|\boldsymbol{y}^*\|_2=1$，带入到式(4)中，可得：

$$\begin{aligned}\cos\theta=\frac{(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{y}^*}{\|\boldsymbol{x}^*\|_2\|\boldsymbol{y}^*\|_2}=\frac{1}{2}\sum_{i=1}^{N}2x_i^*y_i^*=\\ \frac{1}{2}\sum_{i=1}^{N}2x_i^*y_i^*-\frac{1}{2}\sum_{i=1}^{N}(x_i^*)^2-\frac{1}{2}\sum_{i=1}^{N}(y_i^*)^2+1=\\ 1-\frac{1}{2}\sum_{i=1}^{N}(x_i^*-y_i^*)^2=1-\frac{1}{2}\|\boldsymbol{x}^*-\boldsymbol{y}^*\|_2^2\end{aligned} \tag{5}$$

去掉式(5)中的平移因子1和比例因子1/2，对于二范数归一化后的价格向量，可以使用负欧几里德距离 $-\|\boldsymbol{x}^*-\boldsymbol{y}^*\|_2$ 作为价格向量的相似性度量函数。

1.3 SVM

目前，有许多股票预测的研究方法，如遗传算法、决策树、马尔科夫链[18]等，这些方法在股票这种非线性、高噪声、波动性较强的数据中，不能很好的预测股票的涨跌，而SVM能利用核函数，通过非线性映射将股票数据映射到一个更高维的空间，利用线性函数对股票的涨跌进行分类。

对于股票数据集 (x_i,y_i)，$i=1,2,\cdots,m$，其中 m 是样本个数，$x_i\in R^n$，$y_i\in\{-1,1\}$，n 表示样本的特征数。

对于单支股票而言，y_i 表示股票的涨跌，-1表示股票下跌，1表示股票上涨。为了验证

AP-SVM算法对于单支股票的预测有影响，簇的标签与单支股票的标签一致。

$$y_i=\begin{cases}-1, \text{down}\\ 1, \quad \text{up}\end{cases} \tag{6}$$

对于线性可分的点，SVM可以通过选择最佳超平面来分类数据。

$$\boldsymbol{w}^{\mathrm{T}}x_i+b=0 \tag{7}$$

如果存在满足方程式(7)的超平面，通过求解以下的优化问题，就能够找到最佳的线性分离超平面。

$$\begin{aligned}&\min_{w,b}\frac{1}{2}\|\boldsymbol{w}\|^2\\ &\text{s.t.}\quad y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)\geqslant 1\end{aligned} \tag{8}$$

引入拉格朗日乘子 α_i，用条件极值求解最优分界面，构造拉格朗日函数。

$$L(\boldsymbol{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{w}\|^2-\sum_{i=1}^{N}\alpha_i(y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)-1) \tag{9}$$

利用对偶求解可得：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i x_j)-\sum_{i=1}^{N}\alpha_i\\ &\text{s.t.}\ \sum_{i=1}^{N}\alpha_i y_i=0\quad \alpha_i\geqslant 0, i=1,2,\cdots,N\end{aligned} \tag{10}$$

直接求解式(10)很难，不能做到100%线性可分，可以通过引入松弛变量 ε，允许有股票预测处于分类错误的一侧。

$$\begin{aligned}&\min\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{N}\varepsilon_i\\ &\text{s.t.}\quad y_i(\boldsymbol{w}^{\mathrm{T}}x_i+b)\geqslant 1-\varepsilon_i,\varepsilon_i\geqslant 0\end{aligned} \tag{11}$$

ε_i 是松弛变量，常数 C 是惩罚系数。如果股票数据点分类错误，划分到其他类，C 越大表示越不想放弃这个点，边界就会缩小，错分点会更少，但是过拟合情况会更加严重。

这样可以求出 α_i^*，由 α_i^* 可得 $\boldsymbol{w}^*$ 和 b^*，就能得出分离超平面

$$\begin{cases}\boldsymbol{w}^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i\\ b^*=y_i-\sum_{i=1}^{N}\alpha_i^* y_i(x_i y_i)\end{cases} \tag{12}$$

分类决策函数

$$f(x)=\text{sign}(\sum_{i=1}^{N}\boldsymbol{w}^* x_i+b^*) \tag{13}$$

1.4　AP-SVM算法模型

利用AP算法将具有相同变化规律的股票归为一个簇，簇内股票大多处于相同行业，这些股票之间相关性很强，股票与股票的数据之间产生影响，将簇内股票数据结合在一起构建新的矩阵簇，相比于单支股票拥有更多的信息量，以便于更好地预测股票价格涨跌趋势，矩阵簇拥有行业信息的同时，数据维度也极大的增加了，而SVM能够有效处理股票这种高维非线性数据，从而提出AP-SVM算法，分类流程如下：

输入：A股股票数据

输出：股票涨跌分类

1)选取中信证券股票，利用式(1)对股票特征做相关性分析，选取相关性较小的2个特征作为模型的输入特征。并用式(2)对输入特征做最大最小归一化处理。

2)获取A股10年内停牌较少的股票数据，根据式(3)对其收盘价做二范数规范化，将处理好的数据构建成一个新的矩阵，根据式(4)和(5)利用AP算法对其进行聚类分析。

3)根据式(6)对簇和簇内股票进行样本标签制作。

4)将特征输入到SVM中，利用网格搜索寻找最优的 C 和gamma，通过式(7)~式(13)，求解出 $\boldsymbol{w}^*$ 和 b^*，可以计算出分类决策函数。

如果分类结果与真实结果一致，则表示分类正确，统计分类正确样本占样本总数的比例，就能得到股票的预测准确率。

2　结果与讨论

从聚宽网站上获取A股2008-09-10至2018-09-10这10年间所有股票的日数据信息，包括时间、开盘价、最高价、最低价、收盘价、成交量、成交额。剔除数据严重缺省、退市、股票数据长时间不变的股票，最后所剩的1 561支股票作为本次实验的研究对象。

2.1　聚类结果分析

将这些股票的收盘价通过时间索引，构建成一个1 561×2 433的矩阵，通过二范式的预处理对收盘价进行处理，将股票数据规范化到一个单位的圆内，避免了不同股票价格间差异过大，而对预测结果产生的影响，然后利用AP算法对新矩阵内的股票数据进行聚类分析，将有相同变化趋势的股票划分到同一个簇中。

使用AP算法聚类的部分股票结果如下：

聚簇1：华银电力，上海电力，华电能源，东方能源

聚簇2：海通证券，中信证券，东北证券，西南证券

聚簇3；新钢股份，鞍钢股份，马钢股份，山东

钢铁

聚簇4:北京银行,浦发银行,交通银行,中信银行

从聚类分类结果来看,大部分簇是按照行业来划分的,簇内股票根据行业的发展而有相对应的波动,这与通常的认识一致。

选取证券行业内股票,统计其涨跌天数和行业内股票涨跌差值,图3中的横坐标表示涨跌差值,即一天内簇内股票上涨的股票数与下跌股票数的差值,若上涨股票数为5,下跌股票数为4,则涨跌差值记为1,纵坐标表示10 a内这些股票涨跌差值所出现的天数,结果如图3所示。

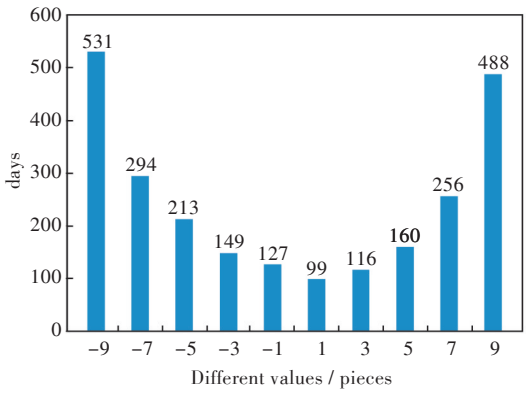

图3 证券行业内股票涨跌天数差

Fig. 3 Days of stock prices in up and down movement in security industry

由图3可以看出,证券行业股票成一个V字型,即2端的数值最大,向中间位置逐渐减少,中间位置的数值最小。图3中的[−1,127]表示证券行业有127 d下跌的股票数比上涨的股票数多1支,[1,99]表示有99 d上涨的股票数比下跌的股票数多1支。证券行业内的股票全涨的天数为488 d,占总交易天数的20%,全跌的天数为531,占总交易天数的21.8%,同涨同跌的天数为1 019 d,占总交易天数的41.8%,超过总交易天数的40%,有1支股票表现和其他股票涨跌不一致的天数为550 d,占总交易天数的22.6%,有2支股票表现和其他股票涨跌不一致的天数为373 d,占总交易天数的15.3%。证券行业内的股票有79.7%的天数有着大致相同的涨跌,可以认为在大的方向上,相同行业内的大部分股票涨跌是相同的。

2.2 AP-SVM预测结果讨论

根据AP算法聚类得到证券行业股票,实验数据为在2008-09-10至2018-09-10期间证券行业的9支股票,每一支有2 433天的交易数据,将其前90%当做训练集,用于训练模型的参数,后10%为测试数据,选择股票前k个交易日的收盘价和成交量作为模型的自变量,模型的因变量是第$(k+1)$个交易日的涨跌,将证券行业内所有股票的特征当作是簇的特征,构建成一个拥有18个特征的矩阵簇。

以往的研究中证实了不同核函数对预测结果有不同的影响,最常用的高斯核函数在大多数情况下都有不错的表现效果,因此本文将在以往的研究基础上,引入高斯核函数,利用网格搜索算法对高斯核函数的参数g和惩罚因子C进行参数寻优。根据SVM基本原理,构建SVM模型对矩阵簇和簇内股票进行实验预测。

本文选择了SVM,BP[19],AP-SVM和AP-BP算法进行了对比实验,在5 d,10 d,15 d,20 d 4种时间间隔的数据进行预测,删除其中停牌日期的数据,对太平洋证券的预测结果如表1所示。

表1 不同时间间隔下四种模型的预测准确率对比表

Tab. 1 Comparison of prediction accuracy of four models at different time intervals

时间间隔 / d	准确率 / %			
	SVM	BP	AP-SVM-9	AP-BP
5	55.4	58.1	58.3	56.8
10	57.3	56.9	59.8	57.9
15	54.8	54.9	56.4	55.8
20	56.7	54.2	51.3	55.6

对比AP-SVM和SVM的预测结果,当时间步长为5 d、10 d和15 d时,AP-SVM的预测准确率要好于SVM,说明聚簇的股票相互作用,会对簇内单支股票的预测准确率有提升作用,当步长慢慢增加,AP-SVM相比单支股票拥有更多行业相关数据,预测准确慢慢变大,但是随着步长继续变大,AP-SVM的输入维度变得很大,样本数据很少,导致准确率慢慢变小。

对比AP-BP和BP的预测结果,当时间步长为10 d、15 d、20 d时,AP-BP的预测效果要好于BP神经网络,说明通过AP聚类和BP神经网络结合,也会提升股票的预测效果。

对比AP-SVM和AP-BP的预测结果,当时间步长为5 d、10 d和15 d时,AP-SVM的预测效果是最好的;当时间步长为20 d时,AP-BP的预测效果要好于AP-SVM。综上所述,当时间间隔取较短时,通过AP聚类和其他算法结合的预测效果会优于单独算法进行预测。

根据AP聚类在一起的股票,大部分是属于同一个行业的,而这些行业之间的股票有着相似的

波动，股票数据之间相互作用提供了单支股票所缺乏的行业信息，从而提升了预测准确率。为了进一步确定簇内多少股票对于预测有作用，利用收盘价计算每两支股票之间相似度，假设2支股票的收盘价矢量表示为 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_N\}$ 和 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_N\}$，则 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 的相似性度量如下：

$$\cos(\theta)=\frac{\sum_{i=1}^{N}x_i y_i}{\sqrt{\sum_{i=1}^{N}x_i^2\times\sum_{i=1}^{N}y_i^2}} \tag{14}$$

利用式(14)计算以收盘价聚类的证券簇内股票之间的相似度，两支股票越相似，它们之间的余弦值越接近1，如图4所示。

	海通证券	中信证券	东北证券	西南证券	太平洋	长江证券	吉林敖东	国金证券	国元证券
海通证券	1	0.978	0.902	0.94	0.905	0.965	0.952	0.934	0.982
中信证券	0.978	1	0.934	0.952	0.942	0.978	0.968	0.946	0.965
东北证券	0.902	0.934	1	0.958	0.964	0.977	0.97	0.927	0.895
西南证券	0.94	0.952	0.958	1	0.956	0.97	0.96	0.942	0.948
太平洋	0.905	0.942	0.964	0.956	1	0.968	0.954	0.957	0.907
长江证券	0.965	0.978	0.977	0.97	0.968	1	0.986	0.96	0.955
吉林敖东	0.952	0.968	0.97	0.96	0.954	0.986	1	0.932	0.947
国金证券	0.934	0.946	0.927	0.942	0.957	0.96	0.932	1	0.92
国元证券	0.982	0.965	0.895	0.948	0.907	0.955	0.947	0.92	1

图4　证券行业内股票的相似度关系图

Fig. 4　Similarity relation schema of stocks in security industry

由图4可知，与太平洋证券走势相似的股票依次是长江证券、东北证券、国金证券、西南证券、吉林敖东、中信证券、国元证券和海通证券。根据与太平洋证券相关性的强弱，依次选取1~9支股票构建AP-SVM算法对太平洋证券进行涨跌预测，选用时间步长为10进行预测的实验结果如图5所示。

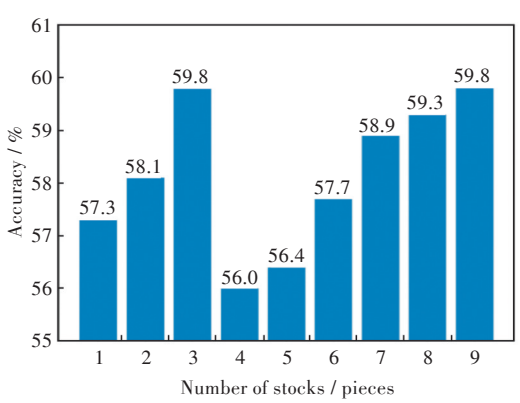

图5　不同股票数目的预测准确率对比图

Fig. 5　Comparison diagram of prediction accuracy for different stock numbers

由图5可知，当选用2支与太平洋证券走势相近的股票时，就能达到用整个簇进行预测的结果，因此可以用簇内的3支股票来预测太平洋证券在不同时间间隔上的表现，预测结果如表2所示。

表2　不同时间间隔下两个不同股票数模型的预测准确率对比表

Tab. 2　Comparison of prediction accuracy of two stock number models at different time intervals

时间间隔 / d	准确率 / %	
	AP-SVM-3	AP-SVM-9
5	55.0	58.3
10	59.8	59.8
15	56.4	56.4
20	53.3	51.3

由表2可知，当时间间隔取10 d和15 d时，选择簇内3支股票和簇内所有股票的预测结果是一样的，说明只要选择用3支走势相近的股票代替整个簇进行涨跌趋势预测，能够减少计算量，缩短计算时间。

3　结　语

本文提出了AP-SVM模型对A股2008~2018年间的股票进行涨跌预测，通过AP算法对A股股票进行聚类，发现聚类在一起的股票是按行业划分的，簇内的股票基本上遵循着同涨同跌的变化规律。通过相关性分析，筛选出相关性小的2个特征，将簇内股票按时间索引构建矩阵簇，对比SVM、BP、AP-SVM和AP-BP在不同时间间隔上对太平洋证券进行预测，结果表明随着时间间隔增大，AP-SVM的预测效果要优于其他3种算法，当时间间隔取最大时，由于AP-SVM的维度变得很大，导致最后的预测结果反而变差。进一步分析簇内股票数目对预测结果的影响，发现在时间间隔取10 d和15 d时，走势最相似的3支股票的预测效果跟整个簇的预测准确率是相同的，可以代替用整个簇预测。

验证了AP算法和其他算法结合，相比单独使用SVM和BP算法，对于股票的准确率预测有提升效果。本文只用同一个簇内有很强相关性的股票进行实验，而不同簇之间也有相关性，挖掘出不同行业涨跌的先后规律，结合深度学习算法是今后的研究方向。

参考文献

[1]　彭丽芳，孟志青，姜华，等. 基于时间序列的支持向

量机在股票预测中的应用[J]. 计算技术与自动化，2006，25(3)：88-91.
[2] 金桃，岳敏，穆进超，等. 基于SVM的多变量股市时间序列预测研究[J]. 计算机应用与软件，2010，27(6)：191-194.
[3] 程昌品，陈强，姜永生. 基于ARIMA-SVM组合模型的股票价格预测[J]. 计算机仿真，2012(6)：343-346.
[4] 李君昌，樊重俊，杨云鹏，等. 基于蒙特卡洛小波去噪的股票投资组合风险优化研究[J]. 计算机应用研究，2018，35(10)：73-77，154.
[5] 黄同愿，陈芳芳. 基于SVM股票价格预测的核函数应用研究[J]. 重庆理工大学学报(自然科学)，2016，30(2)：89-94.
[6] 卢钰. 基于参数优化的支持向量机股票市场趋势预测[D]. 杭州：浙江工商大学，2013.
[7] 孔伟，张彦铎. 基于遗传算法的自主机器人避障方法研究[J]. 武汉工程大学学报，2008，30(3)：110-113
[8] 王曙燕，杨悦，孙家泽. 基于改进粒子群算法的变异体选择优化[J]. 计算机应用研究，2017，34(3)：752-755.
[9] 张贵生，张信东. 基于近邻互信息的SVM-GARCH股票价格预测模型研究[J]. 中国管理科学，2016，24(9)：11-20.
[10] 李辉，赵玉涵. 基于DFS-BPSO-SVM的股票趋势预测方法[J]. 软件导刊，2017(12)：147-151.
[11] 张伟，李泓仪，兰书梅，等. GA-SVM对上证综指走势的预测研究[J]. 东北师大学报(自然科学版)，2012，44(1)：55-59.
[12] 胡蓉. 基于多输出支持向量回归算法的股市预测[J]. 云南民族大学学报(自然科学版)，2007，16(3)：189-192.
[13] 曾岫，彭宏，曾振. 基于分形的股票聚类分析实证研究[J]. 计算机应用与软件，2009，26(7)：104-106.
[14] 柯冰，钱省三. 聚类分析和因子分析在股票研究中的应用[J]. 上海理工大学学报，2002，24(4)：371-374.
[15] 吴微，陈维强，刘波. 用BP神经网络预测股票市场涨跌[J]. 大连理工大学学报，2001，41(1)：9-15.
[16] 张胜东，童雄，张翼，等. 基于BP人工神经网络的球磨机钢球配比预测模型[J]. 武汉工程大学学报，2016，38(3)：299-307.
[17] 刘玉兰，刘小平，邹艳妮. 改进的自适应权值核范数最小化去噪算法[J]. 计算机工程与设计，2018，39(1)：212-217.
[18] 彭舰，孙海，陈瑜，等. 基于马尔科夫链的轻轨乘客轨迹预测新算法[J]. 电子科技大学学报，2018，47(5)：82-87.
[19] 王爱平，陶嗣干，王占凤. BP神经网络在股票预测中的应用[J]. 微型机与应用，2010，29(6)：78-79.

本文编辑：陈小平