doi: 10.3969/j.issn.0490-6756.2018.02.011

基于区域标记法的代价敏感支持向量机在股票预测中的研究

秦 璐, 李旭伟
(四川大学计算机学院, 成都 610065)

摘 要: 针对传统股票预测中单点标记法的缺陷,提出了区域标记法,区域标记法可以为训练分类器提供更多有用信息,在一定程度上减轻了类别不平衡的问题,也更能满足实际任务的需求. 同时,构建了一个 RCS-Trader 模型,该模型使用了代价敏感的支持向量机和 F_S 度量进行优化,相比于传统股票预测方法,RCS-Trader 模型的效果更好,投资回报率更高.

关键词: 区域标记法; 股票预测; 支持向量机; 代价敏感
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2018)02-0277-06

A study of cost-sensitive SVM based on region labeling method in stock prediction

QIN Lu, LI Xu-Wei
(College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: In this paper, the region labeling method is proposed for the shortcomings of single point labeling method in traditional stock forecasting. The region labeling method can provide more useful information for training classifier and alleviate the problem of class imbalance to a certain extent, which is also more suitable for practical needs. At the same time, this paper constructs an RCS-Trader model, which uses cost-sensitive support vector machines and F_S measure to optimize. Compared with traditional stock predicting methods, RCS-Trader model works better and has higher return rate of investment.

Keywords: Region labeling method; Stock prediction; SVM; Cost sensitive

1 引 言

股票投资已经成为一项非常重要而且普遍的金融活动,股票时间序列预测也逐渐成为众多学者研究的热门课题,尤其是转折点的预测研究. 目前,股票时间序列的预测方法主要采用分段线性表示(Piecewise Linear Representation,PLR)进行时间序列的预处理[1,2],在每段标出最低点和最高点,即单点标记法,然后采用不同的方法(如回归分析、数据挖掘和机器学习)进行建模以预测股票时间序列的转折点,进而产生交易信号获取投资收益. 例如,Nair 等人对股票时间序列进行聚类从而产生股票交易决策[3];文献[4]提出了进化趋势反转模型来预测股票交易规则;文献[5]针对股票交易决策信号问题构建了动态阈值模型来预测未来的交易信号.

目前现有的方法大多都是针对股票价格趋势的转折点进行研究,但是,股票市场是一个非常复

收稿日期: 2017-06-28
基金项目: 国家自然科学基金(61173099)
作者简介: 秦璐(1993—), 女, 四川成都人, 硕士生, 研究方向为计算金融. E-mail: qinlu316193@163.com
通讯作者: 李旭伟. E-mail: lixuwei@suc.edu.cn

杂的系统,影响股票价格的因素众多[6],比如:宏观经济状况、政治因素、利率或汇率政策等.因此,对股票市场进行精准预测仍然非常困难[7].本文针对股票时间序列的特点,提出了转折区域(底部区域和顶部区域)的概念,对底部区域和顶部区域的交易点进行标记,即为区域标记法.转折区域相比于转折点而言,能够产生更多的交易决策信号.同时,本文使用代价敏感[8]的支持向量机学习股票买卖点出现的规律,并使用 F_S 度量优化模型,以 F_S 度量最大化为目标进行参数寻优工作,解决了样本中类别不平衡的问题,使训练出的预测模型更加符合实际任务需求.在下文中,我们将本文模型称为 RCS-Trader(Region Cost Sensitive-Trader).

2 RCS-Trader 模型的构建

本文提出的 RCS-Trader 模型是一个以区域标记法和代价敏感支持向量机为基础,以股票价格走势发生改变的转折区域为判断依据的股票交易决策预测模型,可以为投资者提供有效的买卖决策支持.

2.1 数据预处理

特征选择是数据预处理的一个重要过程,RCS-Trader 模型使用了 22 个特征作为样本属性,其中,除了股票交易数据中包含的开盘价、最高价、收盘价、最低价和成交量外,还包含 17 个金融技术指标,详见表 1.

表 1 特征列表

Tab. 1 Feature list

开盘价	open	乖离率	BIAS5,BIAS10
最高价	high	指数平滑移动平均线	MACD
收盘价	close	随机指标	K,D,J
最低价	low	能量潮	OBV
成交量	volume	变动率指标	ROC12,ROC15
成交金额	amount	相对强弱指数	RSI6,RSI12,RSI24
均线	MA5,MA10,MA20	威廉指标	WR

获得样本属性后,我们对数据进行了归一化处理,并通过主成分分析(Principal Component Analysis,PCA)进行降维处理[9],最终获得一个 6 维数据.数据预处理的具体过程如图 1 所示.

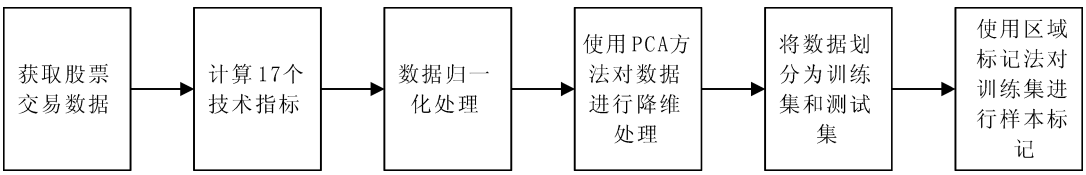

图 1 数据预处理过程

Fig. 1 The process of data preprocessing

2.2 样本标记方法

传统的股票预测模型在标记样本点时均利用 PLR 方法进行单点标记[10-12],即只把 K 线图中某一段的最高点标为卖出点,最低点标为买入点,其他的样本点均标记为不操作(持有或空仓),如图 2 所示.

这种单点标记法,在最低点买入并且在最高点卖出可将收益最大化,但这是一个理想的场景.在实际的股票交易操作中,投资者往往很难把握住精准的最高点和最低点,并且一个波段中可能存在多个最高点或最低点,若只标记单点,则会造成一些重要信息的遗漏和类别严重不平衡的问题,因此,针对上述问题本文提出了区域标记法.

在样本点标记的过程中,本文根据顶部区域和底部区域的概念对样本进行了区域标记,即把 K 线图中某一段顶部区域的样本点全部标记为卖出区域,底部区域的样本点全部标记为买入区域,其他样本点标记为不操作(持有或空仓),如图 3 所示.

图 2 单点标记法

Fig. 2 Single point labeling method

图 3 区域标记法
Fig. 3　Region labeling method

区域标记法可以为训练分类器提供更多有用的信息，从而可以获得效果更好的分类器. 当买入区域中有一个样本预测正确，投资者即可买入；卖出区域中有一个样本预测正确，投资者即可卖出. 由此可见，使用区域标记法获得的收益要比使用单点标记法获得的收益更加稳定，这种标记方法也更符合实际需求.

2.3　代价敏感支持向量机

支持向量机的基本思想是基于训练集在样本空间中找到一个划分超平面，将不同的类别样本分开[13]. 通常，一个标准的支持向量机模型如下：给定训练样本集 $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$，$y_i\in\{-1,+1\}$，划分超平面可用如下线性方程描述[14,15].

$$\omega^T x+b=0 \tag{1}$$

其中，ω 是超平面的法向量；b 是偏差值. 支持向量机的目的是找到具有最大间隔的划分超平面，即满足下式.

$$\begin{cases}\min_{\omega,b}\dfrac{1}{2}\|\omega\|^2\\ \text{s.t. } y_i(\omega^T x_i+b)\geqslant 1,i=1,2,\cdots,m\end{cases} \tag{2}$$

为了使支持向量机的容错性更强，本文为支持向量机引入了松弛变量，允许一些样本点到划分超平面的距离不满足标准支持向量机的条件. 虽然这对分类器来说是种损失，但也使划分超平面不必向这些样本点的方向移动. 本文将这些损失作为惩罚因子 C 加入到目标函数中，因此，标准支持向量机的优化问题就变成以下形式：

$$\begin{cases}\min_{\omega,b,\varepsilon i}\dfrac{1}{2}\|\omega\|^2+C\sum\varepsilon_i\\ \text{s.t. } y_i(\omega^T x_i+b)\geqslant 1-\varepsilon_i\\ \varepsilon_i\geqslant 0,i=1,2,\cdots,m\end{cases} \tag{3}$$

式(3)中惩罚因子 C 表示分类器对离群点的重视程度，C 越大，表示分类器对离群点的重视程度越高. 在代价敏感支持向量机中，并非所有的样本点都有对应的松弛变量，只有离群点才有与之对应的松弛变量. 对不同的离群点赋予不同的惩罚因子，这就是代价敏感学习在支持向量机中的一种应用. 本文通过这种方式在一定程度上解决了样本点中类别不平衡的问题，因此，本文股票预测模型的优化问题可表述为以下形式：

$$\begin{cases}\min_{\omega,b,\varepsilon i}1/2\|\omega\|^2+C_0\sum\varepsilon_i+C_1\sum\varepsilon_j+\\ \quad C_2\sum\varepsilon_k\ \text{s.t. } y_i(\omega^T x_i+b)\geqslant 1-\varepsilon_i\\ \varepsilon_i\geqslant 0,i=1,2,\cdots,p\\ \varepsilon_j\geqslant 0,j=p+1,p+2,\cdots,p+q\\ \varepsilon_k\geqslant 0,k=i+q+1,i+q+2,\cdots,p+q+t\end{cases} \tag{4}$$

其中，$i=1,2,\cdots,p$ 表示标记为不操作的样本点，$j=p+1,p+2,\cdots,p+q$ 表示标记为买入的样本点，$k=p+q+1,p+q+2,\cdots,p+q+t$ 表示标记为卖出的样本点，不同的类别有不同的惩罚因子 C. 本文将使用 3 折-交叉验证的方法对代价敏感支持向量机中的惩罚因子 C、核函数的选择以及相关参数等进行寻优.

2.4　RCS-Trader 模型的优化

传统的股票交易决策预测模型都是以准确率最高作为优化目标，但股票交易决策预测是三分类问题，而且具有类别不平衡的特点，同时本文提出的 RCS-Trader 模型使用区域标记法进行样本标记，所以，针对本文研究的问题和特点，提出了 F_S 度量的概念，并且以 F_S 最大化作为模型的优化目标.

在本文中，F_β 度量定义为准确率(precision)和召回率(recall)的加权调和平均值，它赋予召回率的权重是赋予准确率的 β 倍. 其计算公式如下[16].

$$F_\beta=((1+\beta^2)\times\text{precision}\times\text{recall})/(\beta^2\times\text{precision}+\text{recall}) \tag{5}$$

而 F_S 度量是买入类 F_β 度量($F_{\beta 1}$)和卖出类 F_β 度量($F_{\beta 2}$)的调和平均值，计算公式如下：

$$F_S=(2\times F_{\beta 1}F_{\beta 2})/(F_{\beta 1}+F_{\beta 2}) \tag{6}$$

其中，$F_{\beta 1}$ 和 $F_{\beta 2}$ 具有相同的权重.

经过多次实验测试发现，在本文构建的 RCS-Trader 模型中，$\beta=0.3$ 时的效果最好，在后文中，如无特殊说明，默认 $\beta=0.3$.

3 实验与结果分析

上节构建了 RCS-Trader 股票交易预测模型，使之与传统方法中的单点标记法，即 PLR 方法进行比较，PLR 方法是指利用直线插补的方法提取时间序列中的转折点[17]. 实验除了样本标记方法不同，其他条件均相同. 实验使用的交易规则是通过第 t 天收集计算的数据对第 $t+1$ 天的交易决策做出预测，交易决策包括三类：买入、卖出和不操作(持有或空仓)，分类预测的交易决策都将在第 $t+1$ 天开盘时执行. 买入的决策信号每出现一次都将买入固定金额的股票，而卖出决策一旦出现一次，就将当前所持有的股票全部卖出. 在某段时间内股票交易的总投资回报率由式(7)计算.

$$\text{Rate}=\{\sum_{i=1}^{k}[(1-\text{tax}-\text{charge})\times \text{sell}_i-(1+\text{tax})\times \text{buy}_i]/[1+\text{tax})\times \text{buy}_i]\}\times 100\%$$
$$i=1,2,3,\cdots,k \tag{7}$$

其中，tax 是税率；charge 是股票交易中的手续费；k 是交易次数；sell_i 是股票在第 i 次交易的卖出价格；buy_i 是股票在第 i 次交易的买入价格.

本文随机选用了云南白药(股票代码：000538)、长江电力(股票代码：600900)作为研究对象，股票历史交易数据均来源于 TuShare 平台，并计算了 MACD、BIAS、OBV、WR 等共计 17 个技术指标.

3.1 云南白药区域标记法与单点标记法的预测比较

实验 1 云南白药区域标记法与单点标记法的预测比较. 在实验中，我们获取了云南白药(股票代码：000538)2006 年 7 月 3 日至 2017 年 4 月 28 日共计 2500 个历史交易数据，使用 2015 年 11 月 2 日至 2017 年 4 月 28 日的 252 个交易日作为测试集数据.

实验通过参数寻优得到代价敏感支持向量机的最佳参数是选择 rbf 核函数，$C=1.00$，gamma$=1.099$，同时，本文构建单点标记法模型，通过参数寻优，得到支持向量机的最佳参数选择是 rbf 核函数，$C=0.93$，gamma$=0.500$. 对测试集进行预测，实验结果如图 4 所示. 实验投资回报率(Rate)如表 2 所示.

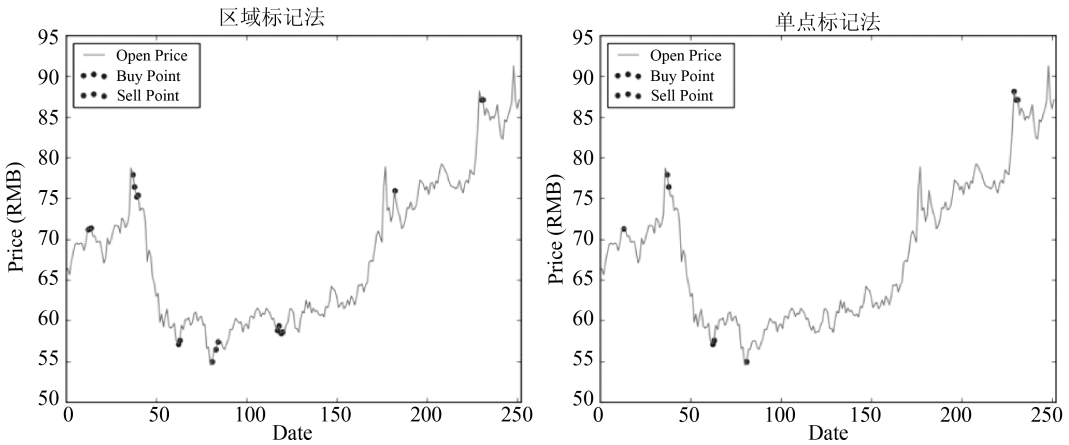

图 4 云南白药区域标记法与单点标记法对比图

Fig. 4 The comparison of region labeling method and single point labeling method of Yunnan baiyao

表 2 云南白药(股票代码：000538)实验结果

Tab. 2 The result of Yunnan baiyao(stock code：000538)

	区域标记法		单点标记法	
买入类预测效果	precision	0.125	precision	0.3333
	recall	0.3333	recall	0.3333
	$F_{\beta 1}$	0.1318	$F_{\beta 1}$	0.3333
卖出类预测效果	precision	0.4286	precision	0.2857
	recall	0.5	recall	0.6667
	$F_{\beta 2}$	0.4337	$F_{\beta 2}$	0.2999
F_S	0.2022		0.3157	
回报率	34.47%		31.77%	

3.2 长江电力区域标记法与单点标记法的预测比较

实验 2 长江电力区域标记法与单点标记法的预测比较. 在实验中，我们获取了长江电力(股票代码：600900)2014 年 4 月 28 日至 2017 年 4 月 24 日共计 622 个历史交易数据，使用 2016 年 11 月 24 日至 2017 年 4 月 24 日的 100 个交易日作为测试集数据.

实验通过参数寻优得到代价敏感支持向量机的最佳参数是选择 rbf 核函数，$C=0.96$，gamma$=1.099$. 对测试集进行测试，同时，本文构建单点标

记法模型，通过参数寻优，得到支持向量机的最佳参数选择是 rbf 核函数，$C=0.93$，gamma$=0.500$，对测试集进行预测实验结果如图 5 所示，实验投资回报率(Rate)如表 3 所示.

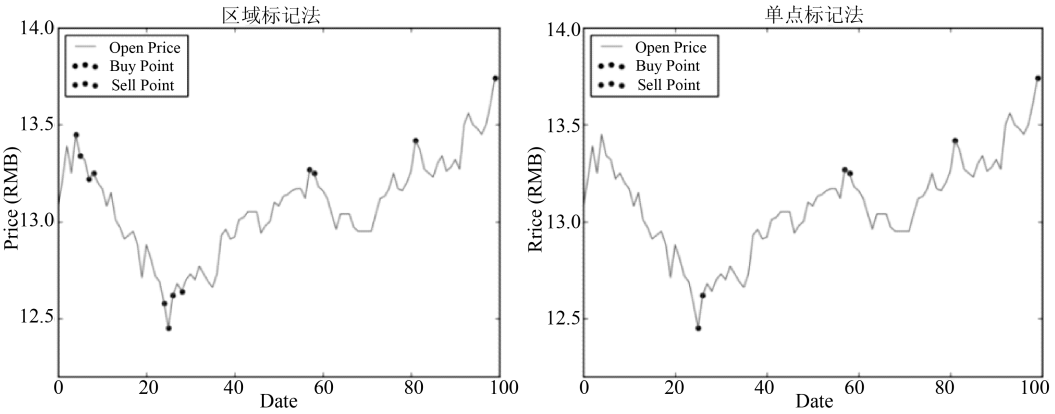

图 5　长江电力区域标记法与单点标记法对比图

Fig. 5　The comparison of region labeling method and single point labeling methodof Yangtze power

表 3　长江电力(股票代码:600900)实验结果

Tab. 3　The result of Yangtze power(stock code:600900)

	区域标记法		单点标记法	
买入类预测效果	precision	0.125	precision	0.3333
	recall	0.3333	recall	0.3333
	$F_{\beta 1}$	0.1318	$F_{\beta 1}$	0.3333
卖出类预测效果	precision	0.4286	precision	0.2857
	recall	0.5	recall	0.6667
	$F_{\beta 2}$	0.4337	$F_{\beta 2}$	0.2999
F_S	0.3554		0.3036	
回报率	5.28%		0	

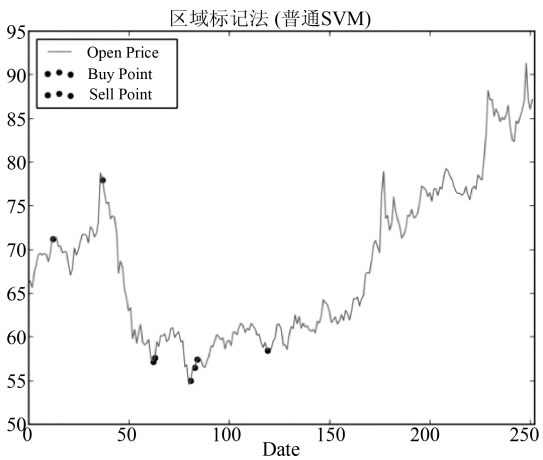

图 6　云南白药普通 SVM 区域标记法

Fig. 6　Yunnan Baiyao Region Labeling Method of essential SVM

3.3　代价敏感支持向量机与普通支持向量机的比较

实验 3　代价敏感支持向量机与普通支持向量机的比较. 在实验中，我们构建了普通支持向量机，利用实验一中云南白药(股票代码：000538) 2015 年 11 月 2 日至 2017 年 4 月 28 日共计 252 个交易日数据进行测试，实验结果如图 6 所示. 实验投资回报率(Rate)如表 4 所示.

表 4　云南白药代价敏感支持向量机与普通支持向量机实验结果对比

Tab. 4　The comparison of Cost-Sensitive SVM and essential SVM of Yunnan baiyao

	代价敏感支持向量机		普通支持向量机	
买入类预测效果	precision	0.125	precision	0.1667
	recall	0.3333	recall	0.3333
	$F_{\beta 1}$	0.1318	$F_{\beta 1}$	0.1738
卖出类预测效果	precision	0.4286	precision	0.5
	recall	0.5	recall	0.1667
	$F_{\beta 2}$	0.4337	$F_{\beta 2}$	0.4291
F_S	0.2022		0.2474	
回报率	34.47%		27.48%	

3.4　实验结果分析

通过对比可以看出，本文提出的 RCS-Trader 模型所使用的基于区域标记法的代价敏感支持向量机的投资回报率均高于单点标记法和普通支持向量机，区域标记法以牺牲一定的召回率(recall)为代价换取了准确率(precision)的提升，这是因为只需在买入区域和卖出区域中预测正确一个或者几个样本点即可，但是单点标记法试图兼顾两者，总体的效果并不如区域标记法，而且单点标记法相比于区域标记法忽略了很多有用信息，例如：某一段最低点周围的较低点也是较为理想的买点，最高点周围的较高点也是很好的卖点，而在单点标记法中，这些重要信息都被忽略了.

4 结 论

针对股票交易的特性,本文在转折点的基础上提出了对样本点的区域标记法,并结合代价敏感支持向量机在一定程度上解决了样本不平衡的问题.同时,通过实验分析可以看出,RCS-Trader 模型也更加容易产生交易决策,更加符合现实操作.

参考文献:

[1] Chang P C, Wu J L, Lin J J. A takagi-sugeno fuzzy model combined with a support vector regression for stock trading forecasting[J]. Appl Soft Comput, 2016, 38: 831.

[2] Chen J, Zhang Y, Zheng J, *et al*. A sliding window mann-kendall method for piecewise linear representation[J]. Sens Transducers, 2014, 175: 187.

[3] Nair B B, Kumar P K S, Sakthivel N R, *et al*. Clustering stock price time series data to generate stock trading recommendations: an empirical study [J]. Expert Syst Appl, 2017, 70: 20.

[4] Zhang X, Hu Y, Xie K, *et al*. An evolutionary trend reversion model for stock trading rule discovery[J]. Knowl-Based Syst, 2015, 79: 27.

[5] Chang P C, Liao T W, Lin J J, *et al*. A dynamic threshold decision system for stock trading signal detection[J]. Appl Soft Comput, 2011, 11: 3998.

[6] Chang T, Chen W Y, Gupta R, *et al*. Are stock prices related to the political uncertainty index in OECD countries Evidence from the bootstrap panel causality test [J]. Economic Systems, 2015, 39: 288.

[7] 陈远, 罗必辉, 蒋维琛, 等. 关于股票价格优化预测的建模仿真研究[J]. 云南大学学报:自然科学版, 2016, 38: 536.

[8] 凌晓峰, SHENG V S. 代价敏感分类器的比较研究[J]. 计算机学报, 2007, 30: 1203.

[9] Li Q, Chen Y, Jiang L L, *et al*. A tensor-based information framework for predicting the stock market [J]. ACM Trans Inform Syst, 2016, 34: 11.

[10] 李丰, 高峰, 寇鹏. 基于分段线性表示和高斯过程分类的股票转折点概率预测[J]. 计算机应用, 2015, 35: 2397.

[11] 钟慧玲, 章梦, 黄维, 等. 基于自适应性窗口的分段线性表示算法[J]. 沈阳工业大学学报, 2014, 36: 79.

[12] Chang P C, Wu J L, Lin J J. A takagi-sugeno fuzzy model combined with a support vector regression for stock trading forecasting[J]. Appl Soft Comput, 2016, 38: 831.

[13] Luo L, You S, Xu Y, *et al*. Improving the integration of piece wise linear representation and weighted support vector machine for stock trading signal prediction[J]. Appl Soft Comput,2017, 56: 199.

[14] 崔伟东, 周志华. 支持向量机研究[J]. 计算机工程与应用, 2001, 37: 58.

[15] 周正松, 李瑶, 陶德元. 支持向量机的凸优化求解[J]. 四川大学学报:自然科学版, 2016, 53: 781.

[16] 王珏, 周志华, 周傲英. 机器学习及其应用[M]. 北京: 清华大学出版社有限公司, 2006.

[17] Shang F H, Sun D C. PLR based on time series tendency turning point [J]. Appl Res Comput, 2010, 27: 2075.