基于多输出学习的沪深300指数预测研究

唐艳琴 潘志松 张艳艳

(解放军理工大学指挥信息系统学院 南京 210007)

摘 要 在股票市场中,人们通常会依赖于股票的历史交易数据来进行推测。目前采用SVM方法进行预测的研究较多,但其模型复杂,耗时较长,而且通常只预测未来1天的数据。文中采用多输出的正则化方法来预测未来多天的走势,对多任务的学习方法进行改进,提出了一种基于多输出的学习方法。实验表明,与SVM支持向量机的方法相比,该方法在沪深300指数预测的均方差值上提高了约10倍,运行时长也减少了近3/4。

关键词 多输出学习,回归,股票预测,任务相关性

中图法分类号 TP391 **文献标识码** A

CSI 300 Index Prediction Research Based on Multi-output Learning

TANG Yan-qin PAN Zhi-song ZHANG Yan-yan

(College of Command Information System, PLA University of Science and Technology, Nanjing 210007, China)

Abstract In the stock market, people usually depend on the historical trading data to predict the trend of future. The method of SVM is more common in the stock prediction, but it is complex and time-consuming, and usually predicts the trend of one day. This article adopted regularization method of multi-output learning to predict the trend of many days. We improved the method of multi-task learning and put forward the method based on multi-output learning. Experiments in CSI 300 index show that the mean square error(MSE) of prediction in this method is about 10 times compared to that in the support vector machine (SVM) method, and the time consuming is also reduced nearly three-quarters.

Keywords Multi-output learning, Regression, Stock prediction, Task correlation

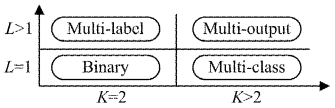
1 简介

在很多实际应用中,我们处理的是多个相关的分类、回归或聚类任务。一个非常简单的方法就是不管任务间的相关性,将这些任务按无相关性来执行。在多任务学习中,将提取出的任务间的相关信息一起考虑。将多个相关任务同时进行学习增加了每个任务的样本量,提高了预测性能。在股票市场中,每天都会产生大量的交易数据,它们之间通常隐藏着大量的有用信息,但一般不容易被挖掘出来。在对股票进行投资时,人们会依赖于股票的历史交易数据进行推测,因此从历史交易数据中挖掘出有用的信息来判断股票未来的走势十分重要。

2 相关工作

在金融市场中,股票数据是一串随着时间变化的数据,因此对其进行训练时也是有时间顺序的,我们希望预测多天的走势情况,因此它的输出结果为一组数据,而这组数据之间具有相关关系,即输出为结构化输出。在很多应用中,分类器的输入是一组观察值,输出是一组类别标记,即 $h\in\mathcal{H}: x=x_1\cdots x_d\in\mathcal{X}\rightarrow y=y_1\cdots y_d\in y$,这类分类器称为结构化机器学习分类器,分类结果称为结构化输出结果。

在结构化输出学习中,输出变量/维数≥2,因此若根据输出变量的值域来划分,则结构化输出可分为多标记(multi-label)学习问题和多输出(multi-output)学习问题。图1给出了不同分类的范例。根据输出变量的值域是否相同,结构化输出可分为多维同质问题和多维异质问题。同质问题表示每维输出的值域相同,语义也相同;而异质问题表示输出变量的值域不尽相同,语义也往往不同,例如输出变量为年龄和性别,年龄的值域为1,2,3,…,性别的值域只有两项:男和女。



注:L 代表输出变量的个数,K 代表输出变量的值域

图1 不同的分类问题范例

在多标记学习中输出的是多个标记,每个标记值仅有两种情况,表示是否是此标记,因此它是一个多维同质问题。在多标记学习中通常会考虑标记间的相关性,经典的多标记学习方法有Ranking和图结构方法[19]。

在多类分类(multi-class)学习中,学习结果只有一个输出变量,值域有多个类,通过训练学习每个类与样本间的关系得到最可能的类别标记,因此它不属于结构化输出范畴。但关于多类分类已经有很多强有力的学习算法,其中典型的学习方法包括multi-class SVM方法[12,14]、神经网络[15]和决策树[16]等。为了研究结构化输出,Thorsten Joachims等人在研究词性标注、自然语言处理等应用时将Markov模型、上下文

本文受国家自然科学基金项目(61473149)资助。

唐艳琴(1977-),女,博士生,副教授,主要研究方向为模式识别,E-mail:nj_tyq@sina.com;潘志松(1973-),男,博士,教授,主要研究方向为模式识别、机器学习;张艳艳(1986-),女,硕士,讲师,主要研究方向为模式识别。

无关语法(CFG)等引入 multi-class SVM 中[13-14]，不仅利用类与样本间的关系，而且根据类与类间的结构(即 Markov 模型、CFG 图)得到多维输出结果，因此“多类学习＋类间的结构化模型”构成了多输出问题。其他经典的多输出学习方法中输出核学习[17-18]、多任务中的正则化学习[3,9-10]方法都能得到多维输出结果，这些均属于多维同质的输出问题。

在多输出学习问题中有一类属于多维异质问题，它的输出变量的语义往往不同，即值域不尽相同。例如在计算广告领域中，计算广告系统需要根据用户在社交网上发布的信息来预测其可能的多个相异特性，包括性别、年龄、职业、性格和政治倾向等。目前这类问题的研究工作相对较少，它既要考虑维内关系，又要考虑维间的关系，如职业可能与政治倾向相关等。

在股票市场中我们希望得到多天的股票走势，对应的是多输出学习问题，每一维输出的值域都是股票指数，属于多维同质问题。

在金融市场中，股票数据是一串随着时间变化的数据，因此对其进行训练时也是有时间顺序的。目前股票价格预测的方法主要有模型法和数据挖掘。模型法是指使用序列进行深层次分析和刻画，如相似性匹配研究[4]、一元时间序列 ARMA、ARIMA 预测和多元时间序列 VARMA 预测[11]等。数据挖掘是指从大量的、不完全的、模糊的、随机的数据中提取隐含的、事前不知但有潜在价值的信息和知识的过程，通常采用的方法有模糊理论、神经网络和支持向量机[5]等。模型法相对简单，但误差较大，因此在对股票进行预测的研究中更多地加入了数据挖掘的方法。

在数据挖掘和机器学习中，分类和回归的通常做法是最小化经验损失加上惩罚项：

$$\min_{w} \mathscr{L}(w)+\Omega(w)$$

其中，w 是训练样本估计的参数值，$\mathscr{L}(w)$ 是指训练集上的经验损失，$\Omega(w)$ 是正则化项。

目前对股票进行预测时多采用支持向量机 SVM 或支持向量回归 SVR 的方法，并且出现了很多改进的算法，如文献[4]引入遗传算法、粒子群算法并采用支持向量机的方法来预测每天的收盘价，但并不能输出多天的值。在预测多天股票值时可采用多输出的方法，首先考虑使用结构化 SVM 方法来解决，即考虑每天股票的升降规律来预测股票未来多天的价值，但在实际应用中股票的升降并没有太大的规律，升降的概率基本一样(为 50%)，因此预测的准确度在 50%～56%之间。我们采用正则化的方法进行学习，同时考虑每天股票之间和任务之间的依赖关系，提出了一种基于多输出的学习方法。

3　基于多输出的股票数据学习

目前对股票数据的研究非常多，但其主要注重于对第二天的预测。我们的研究目标是一次性预测未来多天的走势情况，因此其输出空间大小可达到指数级规模 $|\Omega_y|^d$，$|\Omega_y|$ 表示 y 的可能取值范围，d 为输出空间的维度。基于多任务的思想，将每 p 天的股票上证数据组合成一个任务来预测未来 q 天股票的走势情况。当训练样本数为 m 时，其任务数为 $(m-p)/q$，从日常股票运作中可知股票每天的数据是有依赖关系的，因此假定任务间是有相关关系的，这种关系在下面通过学习得出。由于股票数据是基于时间序列的，因此任务之间的关系是只有前面的任务会对后面的任务产生影响，我们提出了一种新的学习方法来解决基于时序的多任务间的关系。

$$\min_{w,B\geqslant 0} \sum_{i=1}^{t} \| X_i^{\mathrm{T}} W b_i - Y_i \|_F^2 + \lambda_1 \| W \|_1 + \lambda_2 \sum_{i=1}^{t} \| w_i - W b_i \|_F^2 \tag{1}$$

其中，X_i 表示第 i 个任务在 p 天的股票数据，Y_i 为第 i 个任务在第 q 天的收盘价，$W=(w_1,w_2,\cdots,w_t)$，$B=(b_1,b_2,\cdots,b_t)$ 表示第 i 个任务与前面任务间的相关性。因为每个任务只会对后面的任务产生影响，因此 B 为上三角矩阵，下三角可设定为 0，$b_i=(b_{1i},b_{2i},\cdots,b_{ii},0,\cdots,0)^{\mathrm{T}}$。当 B 为单位矩阵 I 时，该方法退化为多任务学习的基本 Lasso 法[2]。当每个任务仅与前一个任务相关而与其他任务都无关时，该方法即为稀疏图正则化的 Lasso 法[3]。

(1)如果知道任务间的相关性，可直接求出 B，那么问题就变成对 W 的求解。通过 $y=x*W*b$ 可求出每天股票的收盘价。

(2)通常情况下，我们知道任务间具有相关性，但无法直接知道相关性的具体数值，这就需要通过计算得出。

采用交替求解法来求解 W 和 B。首先，固定 B 来求解 W。式(1)可变成对式(2)的求解。

$$\min_{w,R} \frac{1}{2} \sum_{i=1}^{T} \| X_i^{\mathrm{T}} W b_i - Y_i \|_F^2 + \lambda_1 \| W \|_1 + \lambda_2 \| WR \|_F^2 \tag{2}$$

其中，$R=I-B$，I 为单位矩阵。

$$W \leftarrow \arg\min_{w} \frac{1}{2} \sum_{i=1}^{T} \| X_i^{\mathrm{T}} W b_i - Y_i \|_F^2 + \lambda_1 \| W \|_1 + \lambda_2 \| WR \|_F^2 \tag{3}$$

在求解式(3)的过程中，存在 1 范数，因此采用加速梯度法 AGM[2]来进行求解。

在求得 W 后，再求解 B。因为此时 W 不变，因此式(1)可等价于求解式(4)。

$$b_i \leftarrow \arg\min_{B} \frac{1}{2} \sum_{i=1}^{T} \| X_i^{\mathrm{T}} W b_i - Y_i \|_F^2 + \lambda_2 \sum_{i=1}^{T} \| w_i - W b_i \|_F^2 \tag{4}$$

这是典型的凸函数求解问题，很多方法都可用来进行求解，此处采用 FISTA[8-9]方法。

具体的算法实现如下：

```
input X,Y; //X:{q*d}*T,Y:{q*1}*T,T 为任务个数
初始化 W 和 B //W:d*T,B:T*T
W_z = W_z_old = W;
B_z = B_z_old = B;
while opts.maxIter //设定最大迭代次数
t_{k+1} = (1+sqrt(1+4t_k^2))/2
W_s = W_z + ((t_k-1)/t_{k+1})(W_z - W_z_old) //固定 B,求解式(3),更新 W
  while true
采用 1 范式投影梯度算法得到 W_zp
    if ||W_zp - W_s||_F^2 < ε
      break;
  end
B_s = B_z + ((t_k-1)/t_{k+1})(B_z - B_z_old)//固定 W,求解式(4),更新 b_i
  while true
采用快速梯度算法得到 B_zp
```

```
    if ‖B_zp−B_s‖_F^2<ε
      break;
  end
W_z_old=W_z;
W_z=W_zp;
B_z_old=B_z;
B_z=B_zp;
  end
end
```

通过计算 $y=x*W*b$ 可求出股票在每天的收盘价，y 为 q 维数据，对应的是 q 天的收盘价。当预测一个新的未知任务(即新的 p 天的股票数据)时，需计算新的任务与已有任务之间的相关性$\tilde{b}$，但这个相关性无法得到，因此我们通过相似性匹配法来得到$\tilde{b}$，进而求得 $y=x*W*\tilde{b}$，x^* 表示新任务。

4 实验

对 2010 年 1 月－2016 年 10 月的上证指数数据进行实验。采用 14 维数据：日开盘价、日最高价、日最低价、日成交量、日换手率、日收盘价、shibor(日级别同业隔夜拆借利率)、PSY(N 日内上涨天数/N * 100)、BTI(广量冲力指标：100 * 上涨家数/(上涨家数＋下跌家数)的 N 日加权移动平均)、ARMS(阿姆氏指标：上涨家数/下跌家数的 N 日异同移动平均)、ABI(绝对幅度指标：上涨家数减去下跌家数所得的差的绝对值)、MCL(麦克连指标)、OBOS(输出超买超卖指标：上涨家数－下跌家数的 N 日异同移动平均 N＝10)和 ADR(涨跌比率：ADR＝N 日上涨家数和/N 日下跌家数和)。

p 表示已知的前 p 天的数据，即 p 天的数据为一组任务；q 表示预测后 q 天的数据，即以前 p 天的数据预测后 q 天的数据。

下面以 p＝9，q＝5 为例来展示本文的处理方法，即以前 9 天的数据预测后 5 天的股票价值。

训练数据共有 1612 个，共产生 320 个任务，每个任务中含有 9 天的数据。因考虑到数据之间有时序关系，数据的顺序在训练及测试中并不会被打乱，每个任务对应一个 5 维的输出结果，即预测的收盘价。

在预测前需对数据进行预处理操作。由于原始数据的属性值之间的差距往往过大，会导致运算效率低及预测不准确等问题，因此首先需对数据进行归一化处理，一般的处理方式是将数据的输入范围控制在[0,1]之间，此时效果最佳。本文采用的归一化方法如下：

$$X_i=\frac{X_i-X_{\min}}{X_{\max}-X_{\min}}$$

其中，$X_{\max}$是样本数据中 X_i 属性的最大值，$X_{\min}$是样本数据中 X_i 属性的最小值。

其次，将数据进行归整化，即将样本数据归整到每个组中。我们将样本数据表示为 $a1,a2,a3,\cdots,a9,a10,\cdots,a14$。$a1-a9$ 表示训练数据，即为 X_i；$a10-a14$ 表示需预测的结果，即为 Y_i。

实验 1　如果不考虑任务间的相关性，即设置 $B=I$

如果仅将训练数据当成测试数据进行测试时，均方误差 MSE＝7.28359e－07；但一旦加入了新的数据再进行预测时，均方误差会增速很快，达到 MSE＝0.00108901。这表明新任务的误差非常大，预测值与实际值间的绝对误差超过 1200 点，相对误差超过了 35%，说明存在过拟合现象。

实验 2　采用交替求解法进行实验

同样仅将训练数据当成测试数据进行测试，其均方误差 MSE＝6.22262e－05，加入新的数据进行预测时，原始数据和预测数据的对比如图 2 所示，绝对误差如图 3 所示。均方误差 MSE＝9.76622e－05，误差相差不大，花费时长为13.30665s。对相同数据使用支持向量机 SVM 的方法得到 MSE＝2.18816e－04，花费时长为 64.322893s。因此，采用基于多任务的方法的性能更优，其 MSE 均方差值提高了约 10 倍，运行时长也减少了近 3/4。

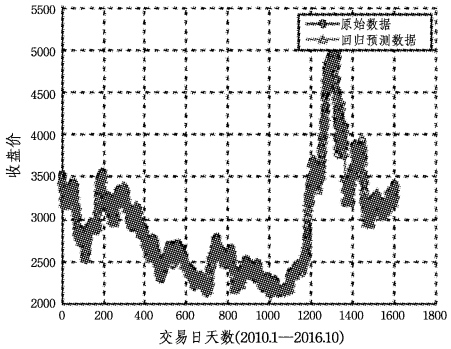


图 2　原始数据和回归预测对比图

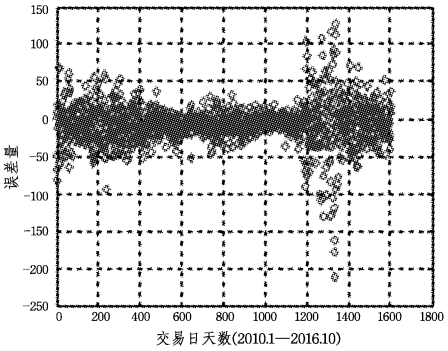


图 3　原始数据和回归预测数据的绝对误差图

鉴于 p,q 的值可调，我们也对 p,q 的不同取值做了测试，得到的均方误差 MSE 如表 1－表 3 所列。

表 1　q＝5 且 p 取不同值时的 MSE 及训练、预测时间

p	q	MSE	Elapsed times/s
30	5	2.32315e－03	49.17108
15	5	8.1267e－04	28.32482
9	5	9.76622e－05	13.30665
7	5	6.53635e－05	20.79470
5	5	9.83589e－05	50.60901

表 2　q＝3 且 p 取不同值时的 MSE 及训练、预测时间

p	q	MSE	Elapsed times/s
15	3	3.66269e－04	89.18031
12	3	1.62219e－04	84.97928
9	3	6.88231e－05	51.42413
5	3	6.12819e－05	50.90407
3	3	5.87862e－05	234.02850

表 3　p＝9 且 q 取不同值时的 MSE 及训练、预测时间

p	q	MSE	Elapsed times/s
9	9	6.45032e－04	3.517991
9	5	9.76622e－05	13.30665
9	3	6.88231e－05	51.42413
9	1	2.84983e－05	1657.44

从表 1 和表 2 中看到，当预测的天数固定时，并不是参与训练的天数越多，精度就越高；相反，参与训练的天数越多，预测的精度越低，这是因为考虑了任务间的相关性后，若每组中训练数据的天数越多，则每个任务之间相似的数据就越少，即任务间的相关性越小，因此就会出现实验 1 的情况，精度会下降很多。

从表 3 中可以看出，当已知天数固定时，预测未来的天数越少，预测的精度越高，需要的时间也就越多。

从图 1 中可以看到，股票上证指数的波动非常大，在预测一组数据时并不是选取的天数越多，预测数据就越准确，间隔几天的数据之间的相关性会大大降低，因此在基于多任务的股票学习中，p 和 q 的取值要根据实际情况来确定。

结束语　股票市场结构复杂，影响因素众多且极不稳定，运用现阶段的手段预测股票价格的变化趋势无疑需要经受各种考验。本文采用多任务的方式进行回归预测研究，并将所提方法与支持向量机 SVM 方法进行了对比，结果表明本方法的精度提高了约 10 倍。但是预测新任务时，很难得到任务间的相关性的确定值，而在多任务方法中，相关性通过学习而得到，因此在后续的研究中，我们将进一步考虑输出间的相关性；同时，在股票预测中，通过回归预测只能得到大致的走向，但实际的升降准确率并不高，即如何提高分类的准确性也是需要进一步研究的内容。

参 考 文 献

[1] BIELZA C, LI G, LARRANAGA P. Multi-dimensional classification with Bayesian networks[J]. International Journal of Approximate Reasoning, 2011, 52: 705-727.

[2] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society, 2011, 73(3): 267-288.

[3] ZHOU J, CHEN J, YE J. MALSAR: Multi-tAsk Learning via StructurAl Regularization[D]. Arizona State University, 2012.

[4] 肖瑞. 不确定性时间序列的降维与相似性匹配研究[D]. 上海：东华大学，2014.

[5] 王琼瑶. 基于改进的支持向量机技术在股票短期价格预测中的应用[D]. 重庆：重庆交通大学，2015.

[6] NG A. Stanford Engineering Everywhere CS229 lecture-Machine Learning[OL]. https://see.stanford.edu/course/cs229.

[7] MURPHY K P. Machine Learning: A Probabilistic Perspective [M]. MIT Press, 2012.

[8] BECK A, TEBOULLE M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. Siam Journal on Imaging Sciences, 2009, 2(1): 183-202.

[9] HAN L, ZHANG Y. Multi-stage multi-task learning with reduced rank[C]// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence(AAAI-16). 2016.

[10] LEE G, YANG E, HWANG S J. Asymmetric multi-task learning based on task relatedness and loss[C]// Proceedings of the 33rd International Conference on Machine Learning. New York, NY, USA, 2016: 230-238.

[11] 吴喜之，刘苗编. 应用时间序列分析：R 软件陪同[M]. 北京：机械工业出版社，2014.

[12] TASKAR B. Learning structured prediction models: A large margin approach[D]. Stanford University, 2004.

[13] JOACHIMS T, HOFMANN T, YUE Y S, et al. Predicting Structured Objects with Support Vector Machines[J]. Communications of the ACM, 2009, 52(11): 97-104.

[14] JOACHIMS T, FINLEY T, YU C N J. Cutting-plane training of structural svms[J]. Machine Learning, 2009, 77(1): 27-59.

[15] HAGAN M T, DEMUTH HB, BEALE M H, et al. Neural network design[M]. PWS Publishing Company Boston, 1996.

[16] MAGERMAN D M. Statistical decision-tree models for parsing [C]// Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics. 1995: 276-283.

[17] DINUZZO F, ONG C S, GEHLER P, et al. Learning output kernels with block coordinate descent[C]// Proceedings of the 28th Annual International Conference on Machine Learning. Bellevue, WA, USA, 2011.

[18] DINUZZO F, FUKUMIZU K. Learning low-rank output kernels [J]. Journal of Machine Learning Research Proceedings Track , 2011(20): 181-196.

[19] ZHANG M L, ZHOU Z H. A Review on Multi-Label Learning Algorithms[J]. IEEE Transactions on Knowledge & Data Engineering, 2014, 26(8): 1819-1837.

（上接第 71 页）

[8] DE C P, O'DWYER M, REILLY R B. Automatic classification of heartbeats using ECG morphology and heartbeat interval features[J]. IEEE Transactions on Bio-medical Engineering, 2004, 51(7): 1196-1206.

[9] 张菲菲，张征. 心电异常自动检测的研究[J]. 电脑知识与技术，2016, 12(4): 197-199.

[10] 李锋，徐涵. 微型便携式心电监护仪的实现[J]. 电子产品世界，2012, 19(12): 52-54.

[11] 程坤. 穿戴式躯感网系统的设计与实现[D]. 上海：东华大学，2015.

[12] 李锋，吴海峰，张能. 异常心电图的自动分析与诊断[J]. 北京生物医学工程，2015, 34(2): 166-174.

[13] 李锋，陈美丽. 一种基于几何特征的 ECG 波形识别算法[J]. 北京生物医学工程，f2015, 34(3): 261-266.

[14] 吴海峰. 心电异常波形智能诊断技术的研究和实现[D]. 上海：东华大学，2014.

[15] 伍育红. 聚类算法综述[J]. 计算机科学，2015, 42(S1): 491-499, 524.

[16] 刘园. 基于划分和层次的聚类算法关键技术研究[D]. 西安：西安理工大学，2014.

[17] 张振亚，程红梅，王进，等. 面向凝聚式层次聚类算法实现的矩阵存储数据结构研究[J]. 计算机科学，2006, 33(1): 14-17.

[18] 任江涛，吴海建，吴向军，等. 一种基于遗传算法的分裂式层次化聚类算法[J]. 计算机应用，2005, 25(11): 2618-2620.

[19] 张峻玮，杨洲. 一种基于改进的层次聚类的协同过滤用户推荐算法研究[J]. 计算机科学，2014, 41(12): 176-178.

[20] 冷莉华，郑智捷. 窦性心律失常心电序列的可视化研究[J]. 计算机科学，2016, 43(S2): 183-185.