

股票趋势预测中 Wrapper 方法的研究与应用

刘利¹, 何先平², 袁文亮¹

(1. 池州学院 数学计算机系, 安徽 池州 247000;

2. 长江大学 信息与数学学院, 湖北 荆州 434023)

摘要:近年来股票市场预测研究一直较受欢迎。大量研究者尝试基于多种数学模型的技术指数及机器学习技术预测股票价格或指数。尽管现有方法展示了较满意的预测成就,但是股票市场是升还是降的预测准确性很少被分析。用 Wrapper 方法从由 23 个技术指标构成的原始特征集中选择最优特征子集,然后用混合不同分类算法的投票法来预测两股票市场的趋势。实验结果表明 Wrapper 方法比常用的 Filter 式特征选择算法如 χ^2 -统计,信息增益,Relief F,对称不确定性,和 CFS 能有更好的性能。此外,提出的投票法超越单一的分类器如 SVM, K 最邻近, BP 神经网络,决策树和 Logistic 回归。

关键词:股票预测; Wrapper; 投票; 特征选择; 分类

中图分类号: TP311

文献标识码: A

文章编号: 1673-629X(2010)01-0213-04

Research and Application of Wrapper Approach to Stock Trend Prediction

LIU Li¹, HE Xian-ping², YUAN Wen-liang¹

(1. Department of Mathematics and Computer, Chizhou College, Chizhou 247000, China;

2. Information and Mathematical College of Yangtze University, Jingzhou 434023, China)

Abstract: The research on the stock market prediction has been more popular in recent years. Numerous researchers tried to predict stock prices or indices based on technical indices with various mathematical models and machine learning techniques. Although these researches exhibit satisfactory prediction accuracy, the prediction accuracy of whether stock market goes or down is seldom analyzed. Employ Wrapper approach to select the optimal feature subset from original feature set composed of 23 technical indices and then use voting scheme that combines different classification algorithms to predict the trend in stock markets. Experimental result shows that Wrapper approach can achieve better performance than the commonly used feature filters, such as χ^2 -statistic, information gain, ReliefF, symmetrical uncertainty and CFS. Moreover, the proposed voting scheme outperforms single classifier such as SVM, kth nearest neighbor, back-propagation neural network, decision tree, and logistic regression.

Key words: stock prediction; Wrapper; voting; feature selection; classification

0 引言

股票市场预测被认为是金融时间序列预测的一项有挑战性的任务。在这一领域有很多用人工神经网络的研究。许多成功的应用显示人工神经网络是时间序列建模和预测的一个非常有用的工具,早期的研究者集中在用人工神经网络预测股票市场,最近的研究趋向杂交好几种人工智能技术。后来提出了遗传算法来

进行特征离散化,人工神经网络连接权的决定来预测股票价格指数,这些方法减少了特征空间的维数,加强了预测性能。

可是,这些研究中有些表明人工神经网络在学习模式上有些缺陷,因为股票市场数据有巨大的噪声和复杂的维数。因此,人工神经网络在噪声数据上展示了不一致和不可预测的性能。然而, BP 神经网络,最流行的神经网络模式,在选择大量的包括相关输入变量、隐层的大小、学习速率和动量常数的控制参数上遇到了困难。

最近,发明了一种新的神经网络算法, SVM。许多传统的神经网络模式落实了实证风险最小化原则,而 SVM 落实了结构风险最小化原则。前者寻求最小化误分类错误或与训练数据的正确解决方案的偏离程

收稿日期: 2009-05-03; 修回日期: 2009-08-02

基金项目: 国家自然科学基金项目(60873021/F0201); 安徽省池州学院院级科研重点项目(XK0902)

作者简介: 刘利(1981-),女,湖北天门人,硕士,讲师,研究方向是概率与数理统计;何先平,硕士,教授,硕士生导师,研究方向是应用数理统计。

度,而后者寻求最小化一个上界泛化误差。此外,SVM的解决方案可能是全局最优的,而其他神经网络模式可能趋向落入局部最优的解决方案。因此,SVM不可能发生过拟合^[1]。

Kim(2003)提出了一种 SVM 方法来预测股票价格的方向。在 Kim(2003)中 11 个技术指标被用做输入量,最好的预测率达到了 59%。为了对付这一挑战,我们尝试用一种合适的特征选择方法从 23 个常用指标中选择最相关的技术指标,然后将选择的技术指标转化成 SVM 分类器来预测两地未来的股票趋势。此外,提出了一种新的投票法,该方法将不同的分类算法与由每个分类器的 Wrapper 方法选择的特征集相结合。

普通的投票法间的不同叫做堆叠,笔者提出的投票法就是普通的堆叠方案仅结合几种不同的分类器来达成共识,在该方法中,进一步用 Wrapper 特征选择算法来为投票法中采用的每一个指定分类找到最好的特征集^[2-5]。

1 Filter 特征选择方法

在许多实际情况下,有太多与股票趋势分类相关的特征了。

从机器学习领域的角度,它们当中有些是不相关的,有些是多余的。人所共知包含不相关的和多余的信息可能引起一些机器学习算法的不正确的结果^[6]。

特征子集选择能被看作通过特征子集空间的一种搜索。在文献中有很多特征选择方法提出来,如:

- (1) χ^2 -统计:这种方法通过计算与类相关的 χ^2 -统计值来测量特征的重要性。
- (2)信息增益:这种方法通过测量与类相关的信息增益来测量特征的重要性。
- (3)对称不确定性:这种方法通过测量与类相关的对称不确定性来测量特征的重要性。
- (4)ReliefF:这种算法是一种对特征互动敏感的特征加权算法。ReliefF 的关键思想是根据它们的值在不同类的例子中区别如何及它们聚类同一类的例子如何来类比特征值。为此,ReliefF 不断地从数据中随机地选择单一的例子,然后找到同类的最近的实例及属于不同类的最近的实例。这些例子的特征值被用来更新每一特征的分。

(5)CFS(Correlation based feature selection):CFS 通过考虑每个特征的个体预测能力及它们中的随机程度来评估特征子集:

$$CFS_s = \frac{k \bar{r}_{cf}}{\sqrt{k + k(k-1) \bar{r}_{ff}}}$$

这里,CFS_s是含有 k 个特征值的一特征子集 s 的分数, \bar{r}_{cf} 是类相关的平均特征($f \in S$), \bar{r}_{ff} 是特征相关的平均特征。一般的 filter 算法和 CFS 间的区别在于当一般的过滤器为每一个特征独立地提供分数时,CFS 给出特征子集的启发式“优点”,并报道它找到的最好的子集。

2 Wrapper 方法加投票机技术

2.1 Wrapper 特征选择算法

Wrapper 方法寻找适于特别算法的最优特征子集,而 Filter 方法尝试测量来自数据集的特征值。Wrapper 方法的概念列在图 1 中,在 Wrapper 方法中,特征子集选择由像一个黑箱的归纳算法来进行。特征子集选择算法用归纳算法自身作为评估函数的一部分来寻找一个好的子集,感应分类器的准确性由准确评估技术来估计,分类算法自身用来决定属性子集。因为 Wrapper 方法在消除特征值时优化分类算法的评估测量,它大多导致比 1 部分描述的所谓的 Filter 方法更大些的准确性。

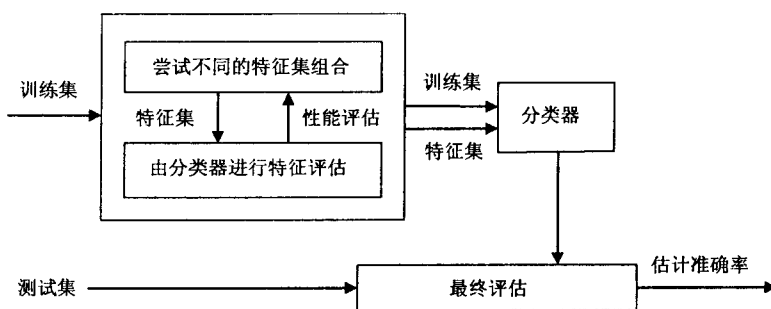


图 1 Wrapper 方法

2.2 投票机技术

投票是人所共知的结合选民的不同意见达成共识的聚集程序,在投票方法的最简单的形式中,每一个数据条目被分给很多票。因为不同的分类算法有各自的优缺点,因此尝试结合 SVM,K 最近邻, BP 神经网络,决策树和 Logistic 回归形成投票法来预测每天的股票价格指数的变化方向^[7]。同时,为不同的分类算法采用由 Wrapper 方法选择的不同的特征值,因为为不同的算法用同一特征集可能是不合适的^[8]。

2.2.1 SVM 支持向量机

SVM 的基本思想是通过内积函数定义的非线性变换将输入空间变换到一个更高维空间,SVM 在这个更高维的空间中找一个线性的超平面,这个超平面在这个空间中具有最大的分类间隔。它的指导原则是同时优化经验风险和模型复杂度,在解决有限样本学习

问题时表现出优异的性能。线性判别函数的一般形式为 $g(x) = w \cdot x + b$, 其相应的分类面为 $w \cdot x + b = 0$ 。SVM 算法对于 2 类线性可分问题, 就是寻找最优超平面, 并且使分类间隔 $(2/\|w\|)$ 最大, 这就相当于使 $\|w\|$ 最小。

2.2.2 k 近邻

KNN 算法是一种基于统计的分类算法。取未知样本 x 的 k 个近邻, 看这 k 个近邻中多数属于哪一类, 就把 x 归于哪一类。具体可以描述为在 N 个已知类别表示的样本中, 找出未知向量 x 的 k 个近邻。设 k_1, k_2, \dots, k_c 分别为待识模式 x 的 k 个最近邻样本实属的样本数。判决准则为: 如果

$$d_m(x) = \max\{d_i(x)\}, i = 1, 2, 3, \dots, c$$

则判 $x \in w_m$ 。其中 $d_i(x)$ 为 w_i 类的判别函数, $d_i(x) = k_i, i = 1, 2, 3, \dots, c$ 。

2.2.3 BP 神经网络

BP 神经网络的层数、传输函数(激活函数)、神经元个数根据网络的训练速度和泛化能力确定^[7,8]。根据理论分析和实践采取如下具体措施: 1. 传输函数; 2. 目标值; 3. 初始化; 4. 层数; 5. 其它参数。

根据上面的分析, BP 神经网络采用双层网络模型。各层次的神经元之间形成全互连连接, 各层次内的神经元之间没有连接。输入层的节点数取为信用数据指标数的 1.5 倍左右。输出层的节点数为两个。输入层为线性传递函数: $f(x) = x$; 输出层传输函数为双曲正切 sigmoid 函数: $f(x) = (e^x - e^{-x}) / (e^x + e^{-x})$ 。实验表明, 该网络经过 3000 次以上的迭代训练具有较好的稳定性能。

2.2.4 决策树

决策树(Decision Tree)是用于分类和预测的主要技术, 它着眼于从一组无规则的事例推理出决策树表示形式的分类规则, 采用自顶向下的递归方式, 在决策树的内部节点进行属性值的比较, 并根据不同属性判断从该节点向下分支, 在决策树的叶节点得到结论。因此, 从根节点到叶节点就对应着一条合理规则, 整棵树就对应着一组表达式规则。基于决策树算法的一个最大的优点是它在学习过程中不需要使用者了解很多背景知识, 只要训练事例能够用属性即结论的方式表达出来, 就能实用该算法进行学习。在该文工作中, 用 C4.5 决策树来预测股票价格的方向变化, 因为 C4.5 决策树在预测应用中执行的很好。

2.2.5 logistic 回归

logistic 回归是因变量是二项分布的一个统计回归模型, 它能被看作一个广义线性模型。模型采用下面的形式:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i}$$

其中 $i = 1, \dots, n$ 。

$$p = \Pr(Y_i = 1 | X) = \frac{e^{\alpha + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i}}}{1 + e^{\alpha + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i}}}$$

这里 p 是属于一个类的概率, $\frac{p}{1-p}$ 是优势率, $\alpha, \beta_1, \dots, \beta_n$ 是回归系数, 回归系数通常由最大似然来估计。

2.3 特征选择和分类

技术分析寻求以不同的方式用不同的指标捕获部分或所有的因素, 旨在分析证券价格的历史行为来决定它的最可能的将来价格。很多技术能用一个精确的数学公式表达。

文中用 23 个技术指标作为总的特征集, 用每天的股票价格指数的变化方向作为预测目标。选择的技术指标如下: OP(开盘价), HP(High Price), LP(Low Price), CP(收盘价), V(成交量), MA6(六日移动平均线), MA12(12 日移动平均线), BIAS6(六日乖离率), BIAS12(12 日乖离率), EMA12(指数平均数指标), EMA26, MACD(平滑异同移动平均线), DIF(差离值), %K, %D, TR(价格波动的真实范围), MTM6(动量指标), MTM12, OSC6(振荡量指标), OSC12, %R5, %R10, OBV(平衡交易量)。

因为该文工作预测每天的股票价格指数的方向, 分别用 1 和 -1 来表示第二天的指数比今天的指数高或低。例子的总数是 365 个交易日, 从 1991 年 8 月跨度到 1992 年 7 月。训练数据的数量是 294, 检验数据是 71。因此近 20% 的数据用来检验, 80% 用来训练。测试数据被用来测试那些没有用来建立模型的数据。首先用 Wrapper 方法来为每个独立分类器找出影响特征, 用投票方案来建立预测模型, 然后用从股票交易公司收集到的数据集来检验模型。

3 实验结果

首先比较 Wrapper 方法和其他的 Filter 特征选择算法, 包括 χ^2 - 统计, 信息增益, ReliefF, 对称不确定性和 CFS 来评估特征选择算法。采用的预测方法是 SVM。接下来, 为了评估所提出的投票法, 比较了投票法与每个单一的分类算法, 包括 SVM, KNN, BP, C4.5DT 和 logistic 回归。Wrapper 方法被用来决定每个单个分类器的特征集。表 1 ~ 表 3 显示了实验结果。

在表 1 中, 给出了用 SVM 分类器加不同的特征选择方法的甲地股票趋势预测准确性的比较。可以看到 Wrapper 方法确实为相应的分类器选择了关键的特征。

表 1 不同的特征选择方法 + SVM 来预测
甲地股票趋势的准确率

特征选择方法	预测准确率
Wrapper	67.61% (46/71)
χ^2 统计	40.85% (29/71)
信息增益	49.30% (35/71)
ReliefF	38.03% (27/71)
对称不确定性	49.30% (35/71)
CFS	40.85% (29/71)

表 2 甲地股票趋势预测的不同分类算法的
预测准确率的比较

分类算法	预测准确率
Wrapper + 投票	74.46% (55/71)
Wrapper + SVM	67.61% (48/71)
Wrapper + KNN	64.79% (46/71)
Wrapper + BP	69.01% (49/71)
Wrapper + C4.5 决策树	64.79% (46/71)
Wrapper + logistic 回归	64.79% (46/71)

表 3 乙地股票趋势预测的不同分类算法的
预测准确率的比较

分类算法	预测准确率
Wrapper + 投票	80.28% (57/71)
Wrapper + SVM	70.42% (50/71)
Wrapper + KNN	64.79% (46/71)
Wrapper + BP	66.2% (47/71)
Wrapper + C4.5 决策树	71.83% (51/71)
Wrapper + logistic 回归	67.61% (48/71)

通过用不同的分类器与 Wrapper 特征选择方法结合的甲地和乙地股票趋势预测的比较分别在表 2 和表 3 中给出。就像期望的那样,所提出的方法达到了最佳表现。这证明在预测算法的帮助下 Wrapper 方法确实能找到最好的特征子集,因为它检测了来自最初特征集的各种子集组合。同时,投票机利用精梳每个分

类成共识,因此比每一个体分类器表现得更好。

4 结束语

文中表明在许多特征选择算法,如 Wrapper, χ^2 -统计,信息增益,ReliefF,对称不确定和 CFS 中,Wrapper 方法能像期望的那样从特征集中找到最相关的特征。实验结果表明投票法加 Wrapper 方法的准确性达到了 80.28% 的准确预测率。同时,实验结果也表明当不同的分类器组合成投票方案时表现更好。在以后的工作中,将尝试不同的分类器组合,如加权投票制,及找到除通用的技术指标外的其他有用的特征,以在股票市场趋势预测应用中达到更好的表现。

参考文献:

- [1] Hastie T, Tibshirani R, Friedman J. 统计学习基础——数据挖掘、推理与预测[M]. 范明,译. 北京:电子工业出版社,2004.
- [2] Cristianini N, Shawe-Taylor J. 支持向量机导论[M]. 北京:电子工业出版社,2004.
- [3] 李蓉,叶世伟,叶忠植. SVM-KNN 分类器——一种提高 SVM 分类精度的新方法[J]. 电子学报,2002,30(5): 745-753.
- [4] 黄超. 基于特征分析的金融时间序列挖掘若干关键问题研究[D]. 上海:复旦大学,2005.
- [5] 邓乃扬,田英杰. 数据挖掘中的新方法——支持向量机[M]. 北京:科学出版社,2004.
- [6] Enke D, Thawornwong S. The use of data mining and neural networks for forecasting stock market returns[J]. Expert systems with applications, 2005,18(29):201-205.
- [7] West D. Neural network credit scoring models[J]. Computers Operation Research,2000,27(11-12):1131-1152.
- [8] Haykin S. 神经网络原理[M]. 叶世伟,史忠植,译. 北京:机械工业出版社,2004.

(上接第 150 页)

用,2008,28(10):2690-2692.

- [5] 张鲁峰,熊志辉,李思昆. 基于虚拟微处理器的嵌入式软件开发与系统验证环境[J]. 计算机研究与发展,2003,40(11):1657-1661.
- [6] Adir A, Almoq E, Fournier L, et al. Genesys-Pro: innovations in test program generation for functional processor verification[J]. Design & Test of Computers, IEEE, 2004,21(2):84-93.
- [7] 王雷,王旭,李巍. 计算机仿真系统生成工具 SIMS 的设计与实现[J]. 系统仿真学报,2005,6(17):1392-1395.

- [8] Cheng K, Krishnakumar A. Automatic generation of functional vectors using the extended finite state machine model[J]. ACM Transactions on Design Automation of Electronic Systems,1996,1(1):57-79.
- [9] 姚英彪,刘鹏,姚庆栋,等. 微处理器功能验证程序生成[J]. 计算机辅助设计与图形学报,2006,18(10):1484-1490.
- [10] 郑德春,姚庆栋,刘鹏,等. 基于软硬件协同仿真平台的功能仿真测试方法[J]. 电路与系统学报,2008,13(2):135-139.