Statistics and Application 统计学与应用, 2016, 5(3), 276-280
Published Online September 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.53027

Hans汉斯

Zinc Futures Price Forecasting Model Based on Data Mining Model

Yongzhong Tian

Yunnan Copper Industry Limited by Share Ltd., Kunming Yunnan
Email: kinki8008@hotmail.com

Received: Aug. 29th, 2016; accepted: Sep. 13th, 2016; published: Sep. 20th, 2016

Abstract

Predicting futures market price reasonably can avoid risks and get benefit. In this paper, we used five kinds of data mining models—support vector machine (SVM) regression, decision trees regression, bagging regression, boosting regression, random forests regression—to predict zinc futures price. It has got good results, whose accuracy rate can reach above 60% for predicting zinc futures price change direction within one month.

Keywords

Zinc Futures, Data Mining Model, Random Forest

基于数据挖掘模型的锌期货价格预测模型

田永忠

云南铜业股份有限公司，云南 昆明
Email: kinki8008@hotmail.com

收稿日期：2016年8月29日；录用日期：2016年9月13日；发布日期：2016年9月20日

摘 要

对期货市场的价格进行合理地预测，可以规避风险，获得收益。本文利用支持向量机(SVM)回归、决策树(RPART)回归、Bagging回归、Boosting回归、随机森林(Random forest)回归五种数据挖掘模型对锌

文章引用: 田永忠. 基于数据挖掘模型的锌期货价格预测模型[J]. 统计学与应用, 2016, 5(3): 276-280.
http://dx.doi.org/10.12677/sa.2016.53027

期货的价格进行预测，预测结果良好，对一个月后锌期货价格变动方向的准确率在60%以上。

关键词

锌期货，数据挖掘模型，随机森林

1. 引言

期货市场作为现货市场的重要组成部分，对现货市场的价格具有指导和发现功能，对期货市场的价格进行预测，不仅能够通过套期保值的方式规避风险，还能利用对期货价格的预测，进行投机套利活动，获取经济收益。因此，本文利用五种数据挖掘模型，利用锌期货的基本面数据、行业数据以及国内外宏观经济数据，对未来一个月、两个月、三个月、六个月期货锌价格的涨跌方向进行预测。

数据挖掘算法诞生于 20 世纪末，是一种全新的智能分析技术和方法，具有非常广阔的应用和发展前景。自诞生以来，数据挖掘算法就受到了理论界和实务界的广泛关注，它指的是通过数据挖掘技术从大量数据中提取出隐含在其中的规律，通过对一系列训练集中的样本数据进行分析，生成一定的规则，并将这规则进行推广运用于测试及过程中。

Mcclean， Scotney 和 Shapcott (2002) [1]将遗传算法引入到分析模型中，对指数化的投资组合进行优化，得到了一定的效果；Hassan 和 Nath (2007) [2]提出了一种基于机器学习的股票价格预测方法，采用了连续的隐马尔科夫模型，将股票的日开盘价，最高价，最低价与收盘价作为模型输入，预测股票的未来日收盘价；杨国梁、赵社涛、蓝柏雄(2001) [3]利用证券分析技术和随机游走模型对中国金融市场进行分析，得出我国金融市场的证券收益存在可预测成分的结论；冯建、邱菀华(2012) [4]用一种基于信息熵的神经网络数据分类方法，通过所有神经元的统计权重信息对输入数据进行投票分类。结合关于数据挖掘算法的文献回顾可以发现，近年来国内外学者在金融分析中，对数据挖掘算法的重视程度越来越高，而数据挖掘算法也因为自身的各种优势越来越被大量的金融投资者所接受。

本文选用的数据包含锌的基本面数据、行业数据以及国内外宏观经济数据三个方面，具体数据情况如表 1 所示。

本文所采用的五种数据挖掘模型分别是支持向量机(SVM)回归、决策树(RPART)回归、Bagging 回归、Boosting 回归、随机森林(Random forest)回归。

2. 理论方法综述

支持向量回归模型的思路为：构建一个超平面 $f(x)=w\cdot\tau(x)+b$，其中，$\tau(x)$ 为因变量向高维空间的映射，对支持向量机的回归，需要使得每个观测点 y 与 $f(x)$ 的离差最小。

决策树回归模型就是把决策树的思想用于回归的方法中，其分析的思路主要是通过选择一个变量也就是一个属性，这个变量使得通过这个进行分类后的两部分的分别的残差平方和的和最小。然后在分叉的两个节点处，再利用这样的准则，选择之后的分类属性。一直这样下去，直到生成一颗完整的树。

Bagging 回归模型、Boosting 回归模型、随机森林回归模型都是基于决策树回归的分析思路，Bagging 回归模型利用自助法抽样(Bootstrap)的方式，建立更多的回归树，其预测结果为每棵回归树加权平均。

Boosting 模型与 Bagging 回归模型类似，两者的区别在于 Boosting 回归在自助法抽样时，会根据前一棵回归树的误差，对抽样的权重进行调整，从而对模型进行修正，是的模型的适应能力，进一步提升，模型的拟合效果更加良好。

Table 1. Analysis variables used in the prediction process
表 1. 预测过程中所选用的分析变量

变量类别	变量符号	变量名称
基本面数据	x1	ILZSG:全球精炼锌过剩/缺口:当月值
	x2	库存小计:锌:总计:月
	x3	总库存:LME 锌:月
	x4	产量:锌:当月值
	x5	产量:锌选矿产品含锌量:当月值
	x6	进口数量:精炼锌:当月值
	x7	进口数量:锌矿砂及精矿:当月值
行业数据	x8	产量:镀层板(带):当月值
	x9	产量:汽车:当月值
	x10	销量:汽车:当月值
	x11	国房景气指数
	x12	70 个大中城市新建住宅价格指数:当月同比
	x13	商品房销售面积:累计值
	x14	房屋新开工面积:累计值
	x15	新房待售面积
	x16	沪伦比
	x17	升贴水
宏观经济数据	x18	产量:发电量:当月值
	x19	中国狭义货币 M1
	x20	中国广义货币 M2
	x21	中国进出口金额:当月值
	x22	中国贸易差额:当月值
	x23	中国制造业 PMI
	x24	中国 CPI
	x25	美国制造业 PMI
	x26	美国核心 CPI
	x27	实际美元指数
	x28	美国 OECD 综合领先指标
	x29	美元兑人民币汇率
	x30	美国:新增非农就业人数:季调(初值)
	x31	美国:失业率:季调

随机森林是由随机放回的样本形成的决策树组成的，其特点是这些决策树的每一节点的分割变量不是由所有的自变量竞争产生的，而是由随机选取的少数变量产生，因此不仅产生的每棵决策树的样本是随机的，每棵树的每个节点的产生也是随机的。这些随机产生的决策树数目很大，因此称为随机森林，

结果的投票(或平均)是等权的。随机森林在每次自助法(Boostrap)放回抽样后，产生一棵决策树，抽样次数等于树的颗数，随机森林分类算法的关键在于在生成每棵树的时候，每个节点的变量都仅仅在随机选出的少量变量中产生。因此，不但样本是随机的，而且每个节点变量的产生都有很大的随机性。随机森林算法会让每棵树尽量生长，并且不会进行修剪。随机森林算法同样也可以避免“过拟合”现象的产生，同时它并不惧怕维数很大的数据，并且能够对没个变量的重要性进行分析，可以广泛的运用与各种高维数据。

3. 模型分析

下面我们利用模型对一个月后期货锌的价格变动方向进行分析预测，预测过程主要分为以下几个步骤：一、分别利用 5 种数据挖掘模型，结合本文所选用的分析变量，利用过去索要预测月份前 12 个月的数据对下一交易月份期货锌的价格变化方向进行预测，分别得出 5 中模型的预测方向；二、在利用过去 15 个月作为测试集，采用滚动预测的方式分别计算出 5 种模型预测的准确率；三、结合 5 中模型不同的预测结果和 5 中模型在测试集中的准确率，计算出综合上涨(下跌)概率。

根据上面的分析结果，我们利用训练期间中的所有数据对 2015 年 12 月锌期货价格相对于 2015 年 11 月的变化情况进行预测，五种模型中，SVM 模型判断为上涨、决策树模型判断为上涨、Bagging 模型判断为下跌、Boosting 模型判断为下跌、Random forest 模型判断为下跌。我们选择的测试区间为 15 个月：2014 年 9 月~2015 年 11 月，经过计算，五种模型在预测集内的准确率分别为：SVM 回归准确率为 60%、决策树回归准确率为 66.67%、Bagging 回归准确率为 60%、Boosting 回归准确率为 73.33%、Random forest 回归准确率为 73.33%。因此，我们认为 2015 年 12 月锌期货价格相对于 2015 年 11 月的综合上涨概率为 20.92%，即为 2015 年 12 月锌期货价格相对于 2015 年 11 月锌期货价格上涨的概率为 20.92%。在分析过程中，我们还可以得到预测过程中变量重要性前五的变量，重要变量指的是在分析过程，对分析结果起重要影响的分析变量，模型分析过程中，重要变量如表 2 所示。

为考察模型在历史行情中的拟合情况，我们利用模型对 2011 年 5 月至 2015 年 11 月的期货锌月度价格的变动方向进行预测，预测结果如图 1 所示，可以发现，模型在 2011 年 5 月至 2015 年 11 月内预测准确率为 64.81%，可以很好的对期货锌的涨跌方向进行预测判断。

4. 总结

可以发现，本文采用的五种模型中，有三种模型认为 2015 年 12 月锌期货的价格较 2015 年 11 月锌期货的价格下跌，结合 5 种模型在测试集中的表现情况，模型最终判断，2015 年 12 月期货锌的价格相对于 2015 年 11 月期货锌的价格上涨的概率为 20.92%，与历史行情相吻合，同时模型在预测过程中所用到的重要变量包括 X2 (库存小计:锌:总计:月)、X8 (产量:镀层板(带):当月值)、X12 (70 个大中城市新建住

Table 2. Important variables in lag one month prediction process
表 2. 滞后一个月进行预测过程中的重要变量

变量序号	变量名称
X2	库存小计:锌:总计:月
X8	产量:镀层板(带):当月值
X12	70 个大中城市新建住宅价格指数:当月同比
X16	沪伦比
X20	中国广义货币 M2

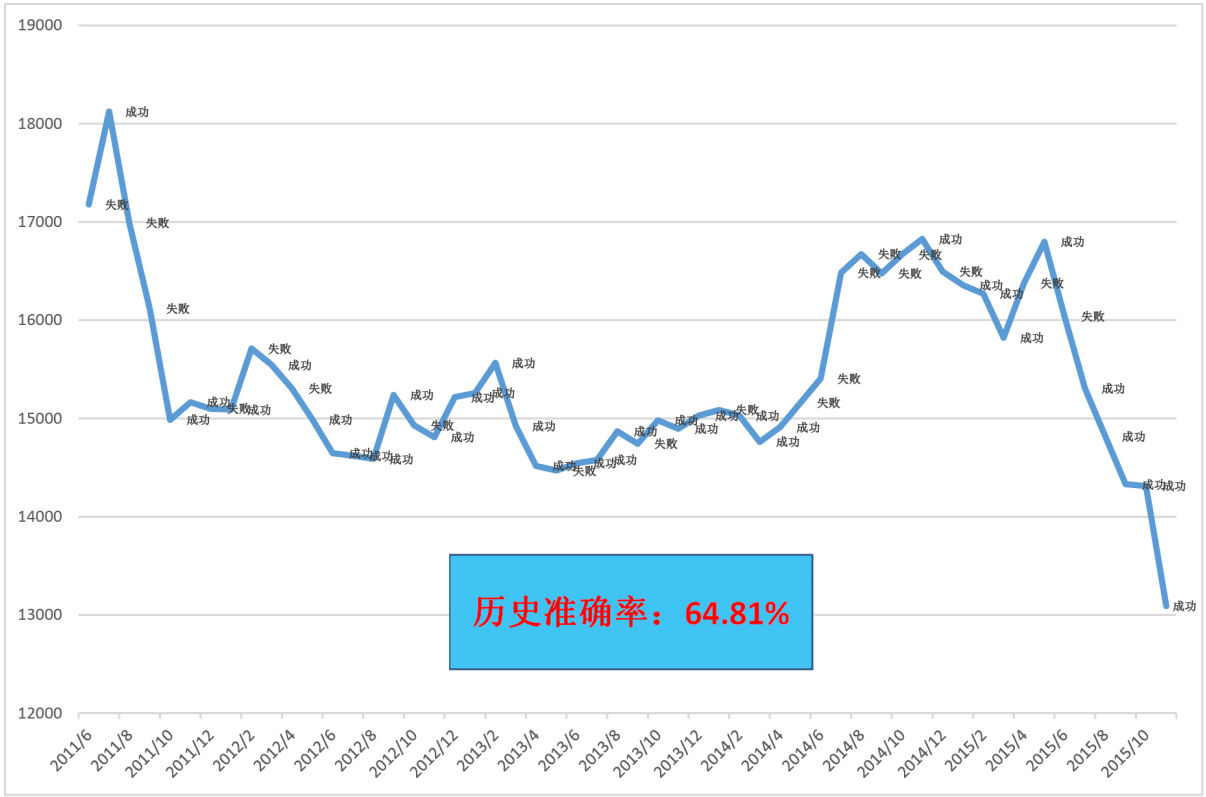

Figure 1. Accuracy of model in lag one month prediction
图 1. 滞后一个月进行预测过程中，模型的准确率情况

宅价格指数:当月同比)、X16 (沪伦比)、X20 (中国广义货币 M2)，均与铝的价格波动要较强的相关性，可以认为这 5 个变量对预测起关键性作用，同时结合历史数据发现，模型在历史行情中表现良好，预测准确率在 60%以上，有较强的参考意义。

参考文献 (References)

[1] Mcclean, S., Scotney, B. and Shapcott, M. (2000) Incorporating Domain Knowledge into Attribute-Oriented Data Mining. *International Journal of Intelligent Systems*, **15**, 535-547. http://dx.doi.org/10.1002/(SICI)1098-111X(200006)15:6<535::AID-INT4>3.0.CO;2-9

[2] Hassan, M.R., Nath, B. and Kirley, M. (2007) A Fusion Model of HMM, ANN and GA for Stock Market Forecasting. *Expert Systems with Applications*, **33**, 171-180. http://dx.doi.org/10.1016/j.eswa.2006.04.007

[3] 杨国梁, 赵社涛, 徐成贤. 基于支持向量机的金融市场指数追踪技术研究[J]. 国际金融研究, 2009(10): 68-72.

[4] 冯建, 邱菀华. 一种基于信息熵的金融数据神经网络分类方法[J]. 控制与决策, 2012, 27(2): 211-215.