Neurocomputing 142 (2014) 48–59

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A causal feature selection algorithm for stock prediction modeling

Xiangzhou Zhang [a,1], Yong Hu [b,*,1], Kang Xie [a], Shouyang Wang [c,d], E.W.T. Ngai [e], Mei Liu [f]

[a] School of Business, Sun Yat-sen University, No. 135, Xingang Xi Road, Guangzhou 510275, PR China
[b] Business Intelligence and Knowledge Discovery, School of Management, Guangdong University of Foreign Studies School of Business, Sun Yat-sen University, Higher Education Mega Center, Guangzhou 510006, PR China
[c] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China
[d] School of Management, Graduate University of Chinese Academy of Sciences, Beijing 100049, PR China
[e] Department of Management and Marketing, The Hong Kong Polytechnic University, Kowloon, Hong Kong, PR China
[f] Department of Computer Science, New Jersey Institute of Technology, University Heights Newark, NJ 07102, USA

ARTICLE INFO

Article history:
Received 16 November 2013
Received in revised form
26 January 2014
Accepted 26 January 2014
Available online 9 May 2014

Keywords:
Stock prediction
Data mining
Feature selection
Causal discovery
V-structure

ABSTRACT

A key issue of quantitative investment (QI) product design is how to select representative features for stock prediction. However, existing stock prediction models adopt feature selection algorithms that rely on correlation analysis. This paper is the first to apply observational data-based causal analysis to stock prediction. Causalities represent direct influences between various stock features (important for stock analysis), while correlations cannot distinguish direct influences from indirect ones. This study proposes the causal feature selection (CFS) algorithm to select more representative features for better stock prediction modeling. CFS first identifies causalities between variables and then, based on the results, generates a feature subset. Based on 13-year data from the Shanghai Stock Exchanges, comparative experiments were conducted between CFS and three well-known feature selection algorithms, namely, principal component analysis (PCA), decision trees (DT; CART), and the least absolute shrinkage and selection operator (LASSO). CFS performs best in terms of accuracy and precision in most cases when combined with each of the seven baseline models, and identifies 18 important consistent features. In conclusion, CFS has considerable potential to improve the development of QI product.

1. Introduction

Quantitative investment (QI) products (models/tools/systems) can provide accurate stock market prediction and help investors significantly alleviate risks of mispricing and irrational trading because of psychological factors, such as overconfidence, mental accounting, loss aversion, and so on [1,2]. One of the key issues of QI product design lies on how to select representative features for prediction. For example, Thawornwong and Enke [3] have adequately demonstrated the effectiveness of recent relevant variables (i.e., representative features) for improving stock direction prediction based on redeveloped probabilistic and feed-forward neural networks. Their work provides evidence of the importance of feature selection for QI product development.

Feature selection, a pre-processing step of data mining, can be used to filter redundant and/or irrelevant features [4]. Feature selection results in simpler model, easier interpretation, and faster induction and structural knowledge [5]. Although many studies have claimed and/or verified that feature selection is the key process in stock prediction modeling [4], identifying more representative features and improving stock prediction are challenging issues that need to be considered. Common feature selection algorithms adopted in stock prediction models include stepwise regression analysis (SRA), principle component analysis (PCA), decision tree (DT), and information gain [3,4,6]. However, all these algorithms can only reveal underlying correlations/associations and cannot determine the direct (i.e., causal) influence of stock features (inputs) on stock return (output).

This paper aims to provide further insight on the application of observational data-based causal discovery approach on feature selection. Based on Pearl's theory [7] and Hu et al.'s study [8], this study proposes the causal feature selection (CFS) algorithm, with the goal of selecting the optimal feature subset for better stock prediction performance and identifying more representative features for better stock market analysis. This study is highlighted on the following two aspects:

First, causal analysis is applied to identify direct influences between variables. Correlation does not imply causation, while causation requires additional counterfactual dependence.

* Corresponding author.
E-mail addresses: zhxzhou@mail2.sysu.edu.cn (X. Zhang), henryhu200211@163.com (Y. Hu), mnsxk@mail.sysu.edu.cn (K. Xie), swang@iss.ac.cn (S. Wang), eric.ngai@polyu.edu.hk (E.W.T. Ngai), mei.liu@njit.edu (M. Liu).
[1] Co-first authors.

http://dx.doi.org/10.1016/j.neucom.2014.01.057
0925-2312/

Causal influences are more consistent over time, which is more attractive to stock investors.

Second, to verify the proposed algorithm more objectively, extensive experiments are conducted for multi-aspect performance comparison. These experiments involve seven baseline prediction models, three popular feature selection algorithms, and various performance measures [accuracy, precision, Sharpe ratio, Sortino ratio, information ratio, and maximum drawdown (MDD)].

To evaluate the proposed CFS, listed companies in the Shanghai Stock Exchanges of China are selected as back-testing subjects, and experimental data cover the period from 2000 to 2012. Experimental results show that CFS performs best in terms of accuracy and precision in most cases and identifies 18 representative features consistently over the entire testing period. Moreover, the constructed prediction models can obtain satisfying and stable investment returns. In conclusion, CFS has considerable potential to improve the performance of existing QI products.

The remainder of this paper is organized as follows: Section 2 reviews common filter-based feature selection algorithms and stock prediction models, and then compares related works in terms of datasets, prediction models, feature selection algorithms, and so on. Section 3 introduces the proposed feature selection algorithm based on the causal discovery algorithm. Section 4 presents the experiment setting of dataset, variables, sliding window test, and evaluation strategies, and then reports the empirical results. Section 5 provides a comprehensive conclusion.

2. Literature review

2.1. Stock prediction

Although the efficient market hypothesis [9] is against stock prediction based on past publicly available information, considerable studies suggest that some markets, especially the emerging markets, are not fully efficient, and prediction of future stock prices/returns may produce better results than random selection [10,11].

Recent studies on stock prediction can be roughly grouped into two types: (a) time series forecasting [12–16] and (b) trend prediction [4,17–20]. A time series forecasting model is trained to fit the historical return/price series of individual stock and is used to predict the future return/price. A trend prediction model is trained to obtain the relationship between various fundamental and/or technical variables and the (rise and decline) movement of stock price (i.e., positive or negative return).

Many popular data mining algorithms have been widely used in stock trend prediction models, including logistic regression (LR) [3,6,21], neural network (NN) [4,21,22], support vector machine (SVM) [23], and decision tree (DT) [21,24]. However, the Bayesian network (BN) [including its variants and naïve Bayes (NB)] has seldom been used directly for stock forecasting/prediction; only Zuo and Kita [25] used BN according to our uncomprehensive search.

Table 1 lists related works in terms of their datasets, prediction models, and feature selection algorithms. Common feature selection algorithms used in stock prediction/forecasting models include SRA, PCA, genetic algorithm (GA), information gain, and so on. Numerous related studies consider both technical and fundamental variables (including economic variables). However, the number of input features used in these studies is different. Currently, no generally agreed-upon representative features for stock prediction are available, and no "best" feature selection algorithm exists. This information motivated us to explore a novel feature selection algorithm by introducing the technique of observational data-based causal discovery to identify a more representative and compact feature set for constructing a simple prediction model with excellent performance.

2.2. Feature selection

Common feature selection algorithms can be grouped into following two types: (1) filter and (2) wrapper approaches [28]. The filter approaches use the general characteristics of the training

Table 1
Comparison of related work.

Work	Dataset	Prediction model	Input features	Feature selection
Chang et al. [26]	S&P500 index and 5 stocks in S&P500	Case based FDT[a]	8 Technical indices	SRA[b]
Tsai et al. [21]	Electronic industry of the Taiwan stock market	Classifier ensembles	19 Financial ratios and 11 economic indicators	–
Tsai and Hsiao [4]	Electronic corporations in Taiwan Stock Exchange	BPNN[c]	22 Fundamental indices and 63 macroeconomic indices	PCA,[d] CART,[e] GA[f]
Lai et al. [27]	3 Taiwan Stock Exchange Corporations	*K*-means + GAFDT[g]	7 Technical indices	SRA
Lee [23]	NASDAQ index	SVM[h]	17 Financial and economic variables	F_SSFS[i]
Enke and Thawornwong [6]	S&P 500 index	BPNN, GRNN,[j] PNN,[k] LR[l]	31 Financial and economic variables	Info. gain
Thawornwong and Enke [3]	S&P 500 index	BPNN, GRNN, PNN, LR	31 Financial and economic variables	Info. gain
Lam [22]	364 S&P companies	BPNN	16 Financial variables and 11 macroeconomic variables	–

[a] FDT: fuzzy decision tree.
[b] SRA: step-wise regression analysis.
[c] BPNN: back propagation neural network.
[d] PCA: principle component analysis.
[e] CART: classification and regression tree.
[f] GA: genetic algorithm.
[g] GAFDT: genetic algorithm-based fuzzy decision tree.
[h] SVM: support vector machine.
[i] F_SSFS: *F*-score and supported sequential forward search.
[j] GRNN: generalized regression neural network.
[k] PNN: probabilistic neural network.
[l] LR: linear regression.