Electronic Commerce Research and Applications 14 (2015) 444–455

Contents lists available at ScienceDirect

Electronic Commerce Research and Applications

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

Concept drift mining of portfolio selection factors in stock market

Yong Hu [a,1], Kang Liu [b,1], Xiangzhou Zhang [c,a,1], Kang Xie [c,*], Weiqi Chen [d], Yuran Zeng [b], Mei Liu [e,*]

[a] Big Data Decision Institute, Jinan University, Guangzhou, PR China
[b] School of Management, Guangdong University of Foreign Studies, Guangzhou, PR China
[c] School of Business, Sun Yat-sen University, Guangzhou, PR China
[d] Faculty of Automation, Guangdong University of Technology, Guangzhou, PR China
[e] Department of Internal Medicine, Division of Medical Informatics, University of Kansas Medical Center, Kansas City, KS 66160, USA

ARTICLE INFO

Article history:
Received 25 November 2014
Received in revised form 19 May 2015
Accepted 14 June 2015
Available online 8 July 2015

Keywords:
Concept drift mining
Stock analysis
Cross-sectional analysis
Causal discovery
Modified Additive Noise Model with Conditional Probability Table
China stock market

ABSTRACT

Concept drift is a common phenomenon in stock market that can cause the devaluation of the knowledge learned from cross-sectional analysis as the market changes over time in unforeseen ways. The widely used cross-sectional regression analysis based on expert knowledge has obvious limitations in handling problems that involve concept drift and high-dimensional data. To discover causal relations between portfolio selection factors and stock returns, and identify concept drifts of these relations, we apply a novel causal discovery technique called modified Additive Noise Model with Conditional Probability Table (ANMCPT). In evaluation experiments, we compares ANMCPT to the conventional cross-sectional analysis approach (i.e., Fama–French framework) in mining relationships between portfolio selection factors and stock returns. Results indicate that the factors selected by ANMCPT outperform the factors adopted in most previous cross-sectional researches that followed the Fama–French framework. To the best of our knowledge, this paper is the first to compare causal inference technique with Fama–French framework in concept drift mining of stock portfolio selection factors. Our causal inference-based concept drift mining method provides a new approach to accurate knowledge discovery in stock market.

1. Introduction

Stock market has always been followed closely by investors all over the world; electronic order book trade in principal markets has reached 48 trillion USD in 2012 (World Federation of Exchanges 2013). However, stock trading is accompanied with great risk, and all market participants strive to trade with higher risk-adjusted returns (Atsalakis and Valavanis 2009, Bahra mmirzaee 2010, Bodie et al. 2005, Fama and French 1992, Malkiel and Fama 1970). One of the key challenges to successful stock market investment is the accurate and timely recognition of informative factors that are closely tied to the expected returns and risks of stocks (called effective portfolio selection factors), as well as the (causal) relationships between these factors and the future returns and risks of stocks.

Concept drift/shift is a phenomenon that the underlying data distribution changes over time, space and conditions, making the relationships between variables found in the past inconsistent with the new data (Delany et al. 2005, Tsymbal 2004, Wang et al. 2003, Zliobaite 2009). According to the efficient market theory (Malkiel and Fama 1970) and numerous empirical researches (Chan et al. 2000, Fama and French 1998, Lam 2002, Lucas et al. 2002, Wang and Di Iorio 2007), effective portfolio selection factors in stock market would change over time and markets; therefore, timely identification of this concept drift is a key problem to stock investment.

Existing researches provide an incomplete view of the concept drift issue in stock market due to limitations in analysis approach and factor selection. Cross-sectional analysis is a conventional approach to discover effective portfolio selection factors, and the most famous one is the cross-sectional regression framework presented in (Fama and French 1992) (hereafter Fama–French framework). The Fama–French framework selects factors according to previous researches and examines the effect of those factors on stock returns using long-term data and multivariable cross-sectional regression model. Although the Fama–French framework were used in many studies and have identified numerous effective factors in various markets, several limitations of this

* Corresponding authors at: School of Business, Sun Yat-sen University, Guangzhou, PR China (K. Xie), and Department of Internal Medicine, Division of Medical Informatics, University of Kansas Medical Center, Kansas City, KS 66160, USA (M. Liu).

E-mail addresses: henryhu200211@163.com (Y. Hu), researcher_lk@foxmail.com (K. Liu), zhxzhou86@foxmail.com (X. Zhang), mnsxk@mail.sysu.edu.cn (K. Xie), isscwqcxz@126.com (W. Chen), zengyuran@hotmail.com (Y. Zeng), meiliu@kumc.edu (M. Liu).

[1] Co-first authors.

http://dx.doi.org/10.1016/j.elerap.2015.06.002
1567-4223/

approach were observed: First, it is ineffective in handling high dimensional data, because testing all possible regression models of candidate factors is often complicated and time consuming, and thus sufficient priori knowledge is needed for feature selection. Second, the selection of factors is theoretically not optimal because most fundamental researches were about US market before 1990; applying factors considered in these researches to other markets might suffer from a concept drift problem. Lastly, it might neglect the nonlinear effects between factors, and, most importantly, regression model identifies correlations rather than causalities.

Compared to the conventional cross-sectional analysis approach, combining causal inference and concept drift adaptation techniques is a more promising approach to discover causalities from high dimensional stock market data. Prior researches have proposed causal discovery techniques for two-to-one and one-to-one causality (Hoyer et al. 2008, Hu et al. 2013, Liu et al. 2014, Zhang et al. 2014). However, the number of factors that influence stocks is uncertain. Thus, this paper proposes a novel causal discovery technique called modified Additive Noise Model with Conditional Probability Table (ANMCPT) to address the problem of many-to-one causality discovery.

In this study, we collected cross-sectional data from China stock market, covering a 13-year period of July 1999 to June 2011. To evaluate the validness and effectiveness of ANMCPT, we first applied the conventional Fama–French framework on a low-dimensional data (Dataset I). Both vertical and horizontal drifts of the relations between factors and stock returns and of the relations among factors were observed. Then, we applied ANMCPT to the same dataset (Dataset I). Results showed that ANMCPT can produce consistent result. Finally, we applied ANMCPT to a high-dimensional data (Dataset II, which consists of 53 factors) and conducted a concept drift analysis. Results revealed obvious concept drifts–the most informative factors changed over time. Moreover, the factors selected by ANMCPT outperformed those by Fama–French framework when being used to construct stock prediction models. This demonstrates the importance of applying causal discovery technique for concept drift mining when we conducts cross-sectional analysis.

The contributions of this paper could be concluded as two points. First, in contrast to most existing approaches for cross-sectional analysis such as the Fama–French framework, our method can not only handle many-to-one causality discovery in high-dimensional dynamic stock markets, but also empirically outperform the classic Fama–French framework. Second, we have clearly exhibited and analyzed the concept drift phenomenon of effective portfolio selection factors. To the best of our knowledge, this paper is the first to compare causal inference technique with Fama–French framework in mining concept drifts of effective portfolio selection factors. The proposed method will provide a new approach and framework for accurate knowledge discovery in stock markets.

The remainder of this paper is organized as follows. Section 2 reviews the related works on cross-sectional analysis and concept drift. Section 3 introduces the Fama–French framework and the ANMCPT method. Section 4 describes two datasets used in our experiments while Section 5 presents the comparative empirical results and analyses. The last section provides conclusions.

2. Literature review

2.1. Cross-sectional analysis

Cross-sectional analysis is a classical investment analysis method. It is different from pattern recognition and time series analysis which are used in technical analysis to identify patterns of price movements of a single stock. Cross-sectional analysis devotes to search for factors that can explain the differences of return between various stocks (Fama and French 1992, 1998; Wang and Xu 2004 and Wang and Di Iorio 2007). Researchers have confirmed many well-known effective portfolio selection factors based on the long-term data (always more than ten years) of a wide range of stocks (such as all stocks in the market, mainly in US stock markets), such as earnings-price ratio (E/P), firm size and book-to-market equity (B/P).

Most effective portfolio selection factors examined by former researches are calculated according to the financial statements of a listed company. However, it doesn't mean that investors who buy a stock having a bad operating situation will definitely lose or vice versa. According to the efficient market theory, the effectiveness of markets will influence the effects of different factors on future stock returns (Malkiel and Fama 1970). On one extreme, if a market is a perfect market, all existing information will be rationally and instantly reflected in stock price, then no investors can gain excess return over the market average, and the differences in return between diverse portfolios are only related to their differences in risk (usually measured by Beta). For example, the Capital Asset Pricing Model (CAPM), established by Sharpe (1964), Lintner (1965) and Black (1972), is such a model that describes the relationship between Beta and expected return of stocks. And it suggests that using Beta is sufficient to describe the cross-sectional differences in expected returns. This idea was supported by (Fama and MacBeth 1973).

However, several subsequent empirical studies challenged CAPM and turned to explore other effective factors for explaining the cross-section of expected stock return. For example, Banz (1981) discovered that smaller firms have a higher average risk–adjusted return than the large firms (Size effect). Although CAPM implies that the effect of leverage can be captured by Beta, Bhandari (1988) found that by controlling the beta and firm size, positive relation exists between leverage and average return. Stattman (1980) and Rosenberg et al. (1985) confirmed the positive relation between expected stock return and book-to-market equity (B/P). Basu (1983) claimed that stocks with higher earnings-price ratio (E/P) earn higher average risk-adjusted return when firm size is controlled.

It is reasonable to anticipate that the effects of the above factors may be partly redundant or decrease when considering other factors. For example, Ball (1978) suggested that E/P is a catch-all proxy for unnamed factors in expected returns. Thus, Fama and French (1992) examined the joint effect of these factors in US stock market of the period between 1962 and 1990, and found that Beta does not help explain the cross-section of average stock returns and the combination of size and B/P can (seems able to) capture the effect of leverage and E/P on average stock returns. Later research by Fama and French (1996) has also considered the effect of cash flow/price (C/P). Many researches have followed Fama and French's approach to investigate these factors in other principal markets (Chan et al. 2000, Fama and French 1998, Lam 2002, Lucas et al. 2002, Wang and Di Iorio 2007).

2.2. Concept drift in stock market analysis

Concept drift indicates the situation where the underlying data distribution change over time, space and condition; thus the relationships between variables found in the old data become inconsistent or irrelevant in the new data (Delany et al. 2005, Tsymbal 2004, Wang et al. 2003, Zliobaite 2009). Concept drift problem can be observed in stock markets when comparing results of existing cross-sectional analysis researches (Chan et al. 2000, Fama and French 1998, Lam 2002, Lucas et al. 2002, Wang and Di Iorio 2007)