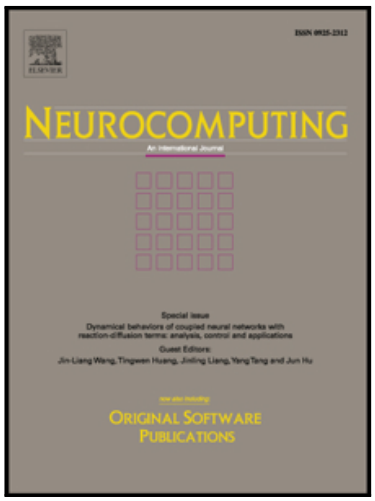

A Comprehensive Cluster and Classification Mining Procedure for Daily Stock Market Return Forecasting

Xiao Zhong , David Enke

PII: S0925-2312(17)31065-2
DOI: 10.1016/j.neucom.2017.06.010
Reference: NEUCOM 18566

To appear in: *Neurocomputing*

Received date: 24 March 2016
Revised date: 13 March 2017
Accepted date: 3 June 2017

Please cite this article as: Xiao Zhong , David Enke , A Comprehensive Cluster and Classification Mining Procedure for Daily Stock Market Return Forecasting, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.06.010

A Comprehensive Cluster and Classification Mining Procedure for Daily Stock Market Return Forecasting

Xiao Zhong and David Enke[1]

Department of Engineering Management and Systems Engineering
Laboratory for Investment and Financial Engineering
Missouri University of Science and Technology
Rolla, MO, USA 65409-0370

Abstract

Data mining and big data analytic techniques are playing an important role in many application fields, including the financial markets. However, only few studies have focused on predicting daily stock market returns, and among these studies, the data mining procedures utilized are either incomplete or inefficient. This paper presents a comprehensive data mining process to forecast the daily direction of the S&P 500 Index ETF (SPY) return based on 60 financial and economical features. The fuzzy *c*-means method (FCM) is initially used to cluster the preprocessed data. A principal component analysis (PCA) is applied next to the entire data set and each of seven clusters. The dimension of the entire cleaned data set is then reduced according to the combining results from the entire data set and each cluster. Corresponding to different levels of the dimensionality reduction, twelve new data sets are generated from the entire cleaned data. Artificial neural networks (ANNs) and logistic regression models are then used with the twelve transformed data sets for classification in order to forecast the daily direction of future market returns and indicate the efficiency of dimensionality reduction with PCA. A group of hypothesis tests are performed over the classification and simulation results to show that the ANNs give significantly higher classification accuracy than logistic regression, and that the trading strategies guided by the comprehensive cluster and classification mining procedure based on PCA and ANNs gain higher risk-adjusted profits than the comparison benchmarks, as well as those strategies guided by the forecasts based on PCA and logistic regression models.

Keywords: Daily stock return forecasting; Data mining; Fuzzy *c*-means (FCM); Principal component analysis (PCA); Artificial neural networks (ANNs); Logistic regression

1. Introduction and methodology

The efficient market hypothesis states that current stock values reflect all available information in the market at that moment, and that the public cannot make successful trades based on that information. However, others believe that the markets are inefficient, in part due to psychological factors of the various market participants, along with the inability of the markets to immediately respond to newly released information [1]. Financial variables, such as stock prices, stock market index values, and the prices of financial derivatives are therefore thought to be predictable. This allows one to gain a return above the market average by examining information released to the general public, with results that are better than random [2].

Stock markets are affected by many highly interrelated factors. These factors include: 1) economic variables, such as interest rates, exchange rates, monetary growth rates, commodity prices, and general economic conditions; 2) industry specific variables, such as growth rates of industrial production and consumer prices; 3) company specific variables, such as changes in company's policies, income statements, and dividend yields; 4) psychological variables of investors, such as investors' expectations and institutional investors' choices; 5) political variables, such as the occurrence and the release of important political events [3], [4]. Each of these factors interacts in a very complex manner [5]. Previous studies have also concluded that the stock market is essentially a dynamic, non-linear, non-

[1] Corresponding author, enke@mst.edu, 573-341-4749, 221 Engineering Management