Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 114 (2017) 465–472

Procedia Computer Science

www.elsevier.com/locate/procedia

Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems, CAS October 30 – November 1, 2017, Chicago, Illinois, USA

Instance Selection Using Genetic Algorithms for an Intelligent Ensemble Trading System

Youngmin Kim[a] and David Enke[b]*

[a]*Laboratory for Investment and Financial Engineering, Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, 205 Engineering Management, 600 W. 14th Street, Rolla, MO, 65409-0370, USA*
[b]*Laboratory for Investment and Financial Engineering, Department of Engineering Management and Systems Engineering, Missouri University of Science and Technology, 221 Engineering Management, 600 W. 14th Street, Rolla, MO, 65409-0370, USA*

Abstract

Instance selection is a way to remove unnecessary data that can adversely affect the prediction model, thereby selecting representative and relevant data from the original data set that is expected to improve predictive performance. Instance selection plays an important role in improving the scalability of data mining algorithms and has also proven to be successful over a wide range of classification problems. However, instance selection using an evolutionary approach, as proposed in this study, is different from previous methods that have focused on improving accuracy performance in the stock market (i.e., Up or Down forecast). In fact, we propose a new approach to instance selection that uses genetic algorithms (GAs) to define a set of target labels that can identify the buying and selling signals and then select instances according to three performance measures of the trading system (i.e., the winning ratio, the payoff ratio, and the profit factor). An intelligent ensemble trading system with instance selection using GAs is then developed for investors in the stock market. An empirical study of the proposed model is conducted using 35 companies from the Dow Jones Industrial Average, the New York Stock Exchange, and the Nasdaq Stock Market from January, 2006 to December, 2016.

Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.

Keywords: Instance selection; Genetic algorithms; Intelligent trading system; Ensemble trading system.

* Corresponding author. Tel.: +1-573-341-4749
E-mail address: enke@mst.edu

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the Complex Adaptive Systems Conference with Theme: Engineering Cyber Physical Systems.
10.1016/j.procs.2017.09.030

1. Introduction

Data mining, or knowledge discovery in databases (KDD), is generally defined as the broad process of discovering hidden valuable knowledge or patterns in large amounts of data [5]. The process is composed of data preprocessing (i.e., data cleaning, integration, transformation, reduction, and discretization), modeling, data analysis, evaluation, and deployment [14]. Data preprocessing is one of the most important data mining processes since the data quality impacts the mining results. If poor data (or information) is coming into a process, unreliable and inconsistent results can be obtained; this is applicable to the concept of garbage in, garbage out (GIGO) in data-driven approaches, such as data mining, pattern recognition, and machine learning.

Data reduction techniques have generally used different approaches, such as feature selection, dimensionality reduction, and instance selection. In the literature, many related studies have shown promising results for data reduction using feature selection and dimensionality reduction when developing intelligent trading system [1, 4, 7, 12, 17]. For example, Zhong and Enke (2017) presented three dimensionality reduction techniques, including principal component analysis (PCA), fuzzy robust principal component analysis (FRPCA), and kernel-based principal component analysis (KPCA) to predict the daily direction of the S&P 500 Index ETF (SPY) return based on 60 financial and economic features. Other studies have also considered the use of technical indicators as features to an intelligent trading system [18, 19]. Chen et al. (2016) proposed a multi-factor time series model based on an adaptive network-based fuzzy inference system (ANFIS) for stock index forecasting. They considered stepwise regression to select technical indicators and then combined this with ANFIS to construct the forecasting model in the Taiwan and Hong Kong stock markets. Kim and Ahn (2012) proposed a new optimization model for artificial neural networks (ANNs) using GAs. It simultaneously optimizes four major architectural factors of ANNs, such as connection weights, the number of neurons in the hidden layer, feature subset selection, and feature transformation (i.e., discretization). Until now, the focus has been on selecting representative features or reducing dimensionality to obtain the same or higher accuracy (i.e., Up or Down prediction), rather than the original data set in stock markets.

However, data reduction carried out without instance selection for classification problems can lead to poor performance since the stock market is a complex system with noisy data. Thus, data preprocessing of instance selection is needed to achieve enhanced performance from learning algorithms [2]. One approach to instance selection is to calculate the distances to neighboring data points using a clustering algorithm (e.g., *k*-nearest neighbors). Another approach to instance selection is to choose the suitable instances (or objects) in the original data set to become the training data set for a learning algorithm. For example, a genetic algorithm (GA) approach to instance selection in ANNs is proposed in [8] to predict stock market movement. This result showed that if noisy and irrelevant instances are eliminated, not only does the classification accuracy increase, but the computational complexity can also be reduced. In particular, a GA is one of the most widely used algorithms for data reduction, such as feature selection, instance selection, and discretization [3, 7, 8, 9].

Nevertheless, from a practical point of view, trading performance is more important than classification accuracy when developing intelligent trading systems for the stock market; for instance, although classification accuracy is higher, trading performance might not be better if the trading system is evaluated using other factors, such as the winning ratio, the profit factor, the payoff ratio, and the number of trades [6, 13, 16]. Therefore, this study proposes a new approach to instance selection that uses GAs to define a set of target labels that can identify the buying and selling signals and then select instances based on the trading measures. Learning algorithms are then trained using the new reduced data set to develop an intelligent trading system. An empirical study of the proposed model is conducted using 35 stocks from the DJIA, NYSE, and Nasdaq Stock Market from January, 2006 to December, 2016. In addition, the results are compared to a conventional approach without instance selection.

2. An intelligent ensemble trading system with instance selection using genetic algorithms

An intelligent ensemble trading system with instance selection using GAs consists of three phases. In the first phase, instance selection using GAs is applied to produce a subset (i.e., instances) from the entire available data set, constructing a new training data set that is used for the learning algorithms. In the second phase, each supervised learning algorithm is trained using the new training data set. In the final phase, the intelligent ensemble trading