# iTrade:
# A Mobile Data-Driven Stock Trading System with Concept Drift Adaptation

*Yong Hu, Institute of Business Intelligence and Knowledge Discovery, Guangdong University of Foreign Studies, Guangzhou, China & School of Business, Sun Yat-sen University, Guanghzhou, China*

*Xiangzhou Zhang, School of Business, Sun Yat-sen University, Guangzhou, China*

*Bin Feng, School of Management, Guangdong University of Foreign Studies, Guangzhou, China*

*Kang Xie, School of Business, Sun Yat-sen University, Guangzhou, China*

*Mei Liu, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, USA*

## ABSTRACT

*Among all investors in the Chinese stock market, more than 95% are non-professional individual investors. These individual investors are in great need of mobile apps that can provide professional and handy trading analysis and decision support everywhere. However, financial data is challenging to analyze because of its large-scale, non-linear and noisy characteristics in a varying stock environment. This paper develops a Mobile Data-Driven Stock Trading System (iTrade), which is a mobile app system based on Client-Server architecture and various data mining techniques. The iTrade is characterized by 1) a data-driven intelligent learning model, which can provide further insight compared to empirical technical analysis, 2) a concept drift adaptation process, which facilitates the model adaptation to market structure changes, and 3) a rigorous benchmark analysis, including the Buy-and-Hold strategy and the strategies of three world-famous master investors (e.g., Warren E. Buffett). Technologies used in iTrade include the Least Absolute Shrinkage and Selection Operator (Lasso) algorithm, Support Vector Machine (SVM) and risk-adjusted portfolio optimization. An application case of iTrade is presented, which is based on a seven-year (2005-2011) back-testing. Evaluation results indicated that iTrade could gain much higher cumulative return compared to the benchmark (Shanghai Composite Index). To the best of our knowledge, this is the first study and mobile app system that emphasizes and investigates the concept drift phenomenon in stock market, as well as the performance comparison between data-driven intelligent model and strategies of master investors.*

*Keywords:* *Adaptive Learning, Concept Drift, Data Mining, iTrade, Mobile Decision Support System, Quantitative Investment*

DOI: 10.4018/ijdwm.2015010104

## 1. INTRODUCTION

Currently, in the Chinese stock market, there are more than 100 million individual investors, among which 95% are non-professional. Institutional investment only accounts for 10% of total market capitalization, and the majority of total outstanding shares are owned by individuals (Wang & Xu, 2004). These individual investors are in great need of mobile decision support system (apps) for trading analysis and decision making.

Although numerous commercial software and freeware, such as *Dazhihui*, *Qianlong*, *Tonghuashun* (from China), *ProfitSource*, *eSingal*, and *VectorVest* (from other countries, see http://stock-software-review.toptenreviews.com) are available to common investors, none of them incorporates data mining functions; they only provide information retrieval, statistical analysis, trade ordering, and technical analysis-based program trading. Their build-in trading rules/models are empirical and static, and thus they cannot adapt to the varying market in time. In addition, they cannot process large-scale financial data due to the limited computing ability of mobile devices (Goh & Taniar, 2004).

Data mining technologies are suitable to address these limitations because they can handle large-scale, non-linear, noisy data. These techniques are used to study data and discover new, hidden, unexpected, valuable trends or patterns from existing databases, and have gained increasing attention in science and business areas (Daly & Taniar, 2004). Client-Server (C/S) architecture can be deployed to partition workloads and share computing resources. Therefore, this paper is to develop a C/S-based Mobile Data-Driven Stock Trading System (iTrade), which can provide intelligent decision support for non-professional investors on mobile devices. The unique characteristics of iTrade can be outlined in three aspects.

First, a data-driven intelligent learning model is constructed for accurate stock (trend) prediction. Compared to the empirical technical trading rule-based stock analysis software, the proposed model is based on a well-known data mining algorithm, Support Vector Machine (SVM) (Vapnik, 1995), which is gaining more and more popularity in stock prediction (Hu et al., 2013; Huang, 2012; Lee, 2009).

Second, a concept drift adaptation process is proposed to identify market structure changes and adapt the learning model to these changes. This is about how to identify the most informative and up-to-date predictors (e.g., fundamental/technical factors) that can explain future excess returns. This process can be realized by combining feature selection and sliding window method (Tsai & Hsiao, 2010; Zhang et al., 2014). In this paper, feature selection is based on the Least Absolute Shrinkage and Selection Operator (Lasso) algorithm because of its effectiveness in sparse and consistent model selection (Zhao & Yu, 2006). Moreover, sliding window method is combined with Lasso to perform adaptive feature selection, which can handle the phenomenon of concept drift that the most informative predictors are ever-changing from time to time, especially from bull to bear market, vice versus. Concept drifts generally exists in stock market and might be derived from mass psychology, macroeconomic, the development of technology, and so on, which makes adaptive feature selection imperative.

Third, a rigorous benchmark analysis is provided to evaluate iTrade, including performance comparisons with 1) the quantified strategies of three world-famous master investors (Warren E. Buffett, William J. O'Neil and Richard Driehaus), and 2) the Buy-and-Hold strategy. This analysis indicates whether the data-driven intelligent model can defeat the human experts.

An application case was carried out to demonstrate and evaluate iTrade. This case was based on a seven-year back-testing using historical stock data from China Stock Exchanges during the period from 2000 to 2011. Comparative evaluations were conducted to examine whether iTrade could gain higher return than that of the Buy-and-Hold strategy on market benchmark (Shanghai Composite Index) and had potential to be a promising alternative to the strategies of master investors.

The rest of this paper is organized as follows: Section 2 is a review of related work. Section 3 outlines the system architecture of iTrade. Section 4 describes the specific design of different modules of iTrade. Section 5 presents the system testing scheme and evaluation results. Section 6 provides a brief conclusion and some directions for future research.

## 2. RELATED WORK

### 2.1. Quantitative Stock Model

Different form the discretionary stock trading strategy, a quantitative stock model applies computerized and systematic techniques and methods to eliminate subjective decisions. In the last few decades, many technologies and methods have been applied to construct stock decision support models, such as mathematical statistics, artificial intelligence, and so on (Ren, Zargham, & Rahimi, 2006). Statistical methods have been used to analyze market behaviors for more than half a century. Data mining is now an emerging technology for stock decision support that gained more and more attentions.

Studies on stock prediction include two types: (a) time series forecasting and (b) trend prediction. A time series forecasting model is trained to predict the future return/price, while a trend prediction model is trained to predict the movement directions (rise or decline) of stock price. Many popular data mining algorithms have been widely used in stock trend prediction models, including logistic regression (Tsai, Lin, Yen, & Chen, 2011), neural network (Enke & Thawornwong, 2005), SVM (Hu et al., 2013; Lee, 2009). Although association rule mining has been well studied (Ashrafi, Taniar, & Smith, 2007; Taniar, Rahayu, Lee, & Daly, 2008), relatively few studies have applied it in stock prediction (Na & Sohn, 2011).

### 2.2. Concept Drift and Adaptive Learning

Concept drift refers to a non-stationary learning problem over time and it indicates that patterns would change in the future, i.e., previous patterns would be invalid and new patterns would emerge (Zliobaite, 2009). In many data mining applications, the most difficult problem is that the concept of interest may be determined by some hidden context (Delany, Cunningham, Tsymbal, & Coyle, 2005), which are hard to uncover by expert judgment/inference. This is about how to timely identify the testing dataset that are no longer consistent with the training dataset (Wang, Fan, Yu, & Han, 2003). The ideal way is to detect the changes as they arise, and re-train the model to adapt to the new data distribution (Kuncheva, 2008).

When using data mining techniques for stock prediction, the problem of concept drifts is inevitable because a stock market is always influenced by mass psychology, macroeconomic, the development of technology, and etc. To address this issue, numerous adaptive learning approaches have been explored to train model dynamically. For example, O, Lee, Lee and Zhang (2006) used moving average to identify different market patterns (*Bearish*, *Bullish*, *Goden Cross* and *Turn Up*) at first, and then trained independent neural network-based prediction models for each pattern (O, Lee, Lee, & Zhang, 2006). Armano, Murru and Roli (2002) assigned each neural network a genetic classifier which is used to identify a potential market trend, and a neural network could take place in the prediction process only if the current market trend enables its genetic classifier (Armano, Murru, & Roli, 2002). Chun and Park (2005) predicted the Korean Stock Price Index by adaptively choosing the optimal model from multiple case-based reasoning predictors according to hit rate (Chun & Park, 2005). Thawornwong and Enke (2004) uncovered the recent relevant variables by using decision tree algorithm, then used neural networks to predict future stock return (Thawornwong & Enke, 2004).

### 2.3. Feature Selection

Tsai and Hsiao (2010) pointed out that there are different factors (i.e. input variables) for

construction of stock prediction models, i.e., no consistent view on the most representative variables exists (Tsai & Hsiao, 2010). Feature selection is a data preprocessing technique that can filter out noisy and unsuitable variables and produce 1) a simpler model, 2) easier interpretation, and 3) faster model induction and structural knowledge (Chen & Cheng, 2009; Zhang et al., 2014).

Common feature selection methods used in stock prediction studies include Step-wise Regression Analysis (SRA) (Chang, Fan, & Lin, 2011; Lai, Fan, Huang, & Chang, 2009), Principle Component Analysis (PCA) (Tsai & Hsiao, 2010; Zhang et al., 2014), Decision Tree (Tsai & Hsiao, 2010; Zhang et al., 2014), Information Gain (Thawornwong & Enke, 2004), and Genetic Algorithm (Tsai & Hsiao, 2010). Recently, more and more researchers have employed Lasso for feature selection in a variety of fields including finance (Hsu, Hung, & Chang, 2008; Wang, Li, & Tsai, 2007). However, to the best of our knowledge, there is only one study by Wang and Tan (2009) that has adopted Lasso-based feature selection for stock prediction (Wang & Tan, 2009). Thus, it is worthy of further investigation of Lasso in stock prediction issue. Detail of the Lasso algorithm is provided in Section 4.1.

## 2.4. Portfolio Optimization

Portfolio optimization is about how to efficiently allocate capital among a set of assets. The foundation of portfolio optimization theory is Markowitz's mean-variance model (Markowitz, 1952). It generates an optimum portfolio based on the idea of minimizing risk and simultaneously maximizing expected returns (Chen, Ohkawa, Mabu, Shimada, & Hirasawa, 2009). The excepted rate of return is quantified by the mean of returns and the risk by the variance of a portfolio which is calculated using the covariance of individual stocks' returns. Investors need to strike a balance between maximizing the expected return and minimizing the risk of investment.

The Markowitz' model can be express as a quadratic programming problem as below:

$$\text{Min } \sigma_{\rho}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij} \tag{1}$$

$$\text{subject to } r_{\rho} = \sum_{i=1}^{n} w_i r_i = r \tag{2}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{3}$$

or:

$$\text{Max } r_p = \sum_{i=1}^{n} w_i r_i \tag{4}$$

$$\text{subject to } \sigma_{\rho}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij} = \sigma \tag{5}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{6}$$

where $n$ is the number of assets; $r_p$ is the expect rate of return; $r$ is the required rate of return; $r_i$ is the expected rate of return of asset $i$; $\sigma_{ij}$ is the return covariance between assets $i$ and $j$, $\sigma_p^2$ is the return variance of the portfolio; and $\sigma$ is the target return variance (risk tolerance). $w_i$ represents the weight of budget invested in asset $i$. In the original Markowitz' model, $w_i \geq 0$, that is, no short selling is allowed. This problem is easy to solve when the required rate of return is specified.

# 3. ARCHITECTURE AND WORKFLOW OF ITRADE

## 3.1. System Architecture

The iTrade is a comprehensive system that incorporates data mining and portfolio optimization technologies and is designed to handle large-scale, non-linear and noisy stock market data. Specifically, it automatically collects the latest stock data, and based on these data it screens out "good" stocks to construct/adjust a portfolio according to user-specified system and trading parameters (e.g., risk aversion parameter, maximum drawdown, and stop loss). Because of the limited computing ability and power supply of mobile devices, it's hard to rapidly execute the tasks of data mining and portfolio optimization in these devices. Therefore, iTrade is developed in a C/S system architecture, as shown in Figure 1.

The technical references of the host server for the system prototype development are 1) Processor: Intel Xeon E5-2403 Processor (1.8GHz, 1066MHz, 10MB, 80W), 2) Memory: 32GB DDR3-1333, and 3) Hard disk: 2TB.

A snap of client interface is illustrated in Figure 2. Users can specify both system and trading parameters in this user interface, and a brief portfolio report would be displayed.

## 3.2. System Workflow

The iTrade server runs in an iterative workflow, as shown in Figure 3.

1. In the data retrieval process, iTrade collects the latest stock data and invokes the data preprocess module to extract proper data for subsequent processing;
2. When iTrade enters a new re-training cycle, the concept drift adaptation process

*Figure 1. The C/S system architecture of iTrade*

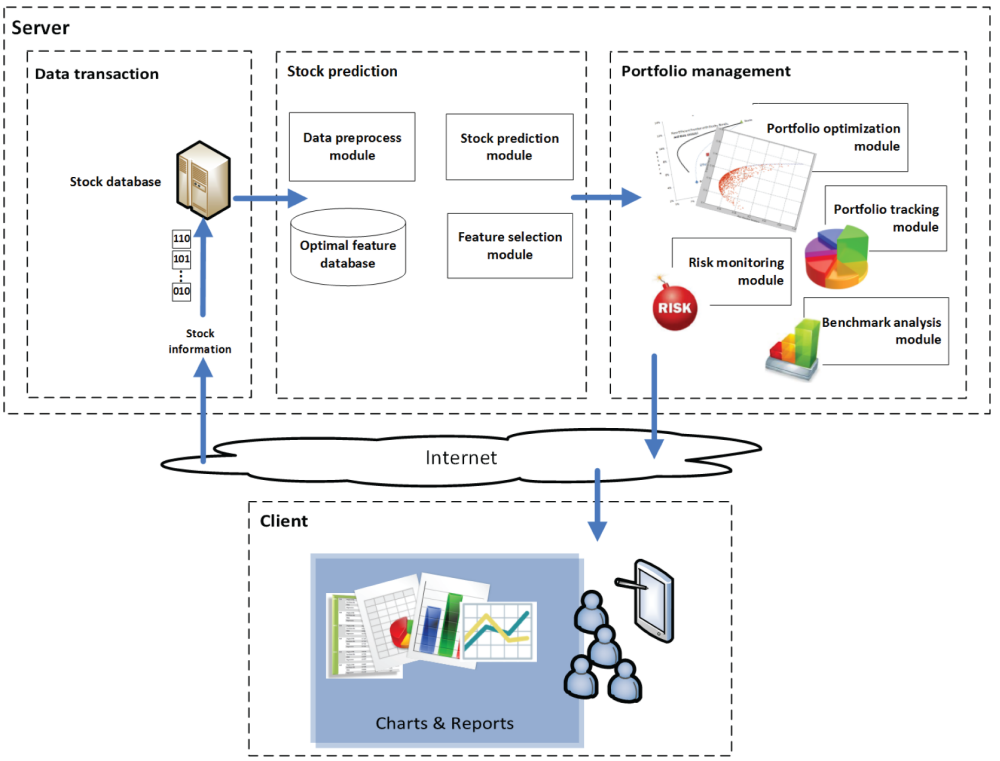
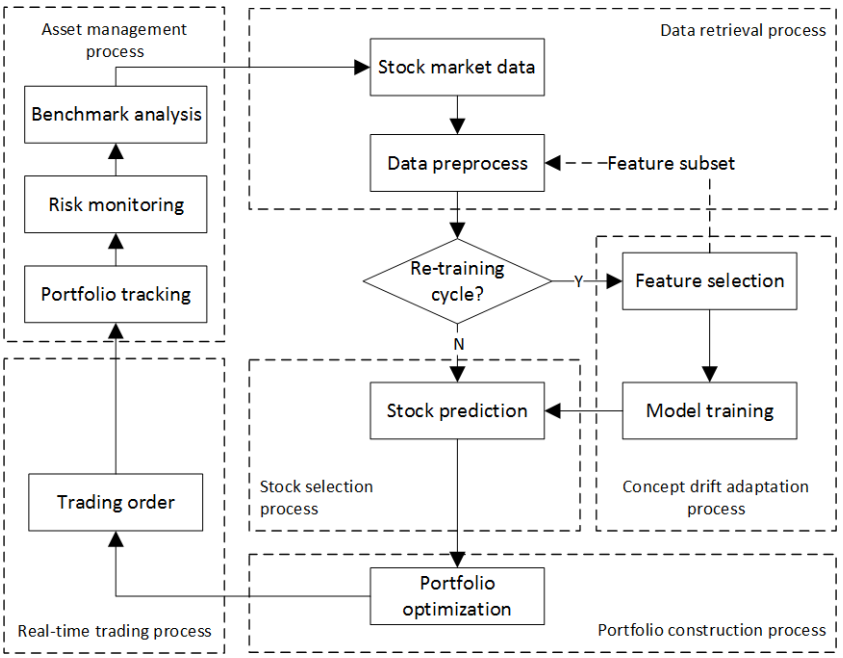


*Figure 2. A snap of the system interface*


*Figure 3. The system workflow of iTrade*



is executed. The feature selection module is invoked to identify the recent informative predictors (i.e., input variables) and forward the optimal feature subset to the data preprocess module for future data extraction. Then, the stock prediction model will be re-trained;

3. For non-re-training cycles, iTrade directly enters the stock selection process, where the stock prediction module is invoked to identify "good" stocks. These stocks are selected as a candidate stock pool for subsequent trading decision;
4. In the portfolio construction process, the portfolio optimization module is invoked to allocate proper capital (weight) among stocks in the candidate stock pool and generate trading signals;
5. In the real-time trading process, iTrade calculates and submits trading orders according to the generated trading signals;
6. In the asset management process, the portfolio tracking, risk monitoring and benchmark analysis modules run in background, continuously monitoring the portfolio status, performing risk management, and providing clients/users with various statistical analysis reports.

## 4. SPECIFIC DESIGN OF ITRADE

This section provides details of several important modules presented in Figure 1.

### 4.1. Feature Selection Module

The Lasso algorithm (Tibshirani, 1996), which is a penalized method for feature selection, is adopted for relevant variable discovery in this study. Suppose there is a data $(\mathbf{x}^{\mathrm{i}}, y_i), i = 1,2,...N$, where $\mathbf{x}^{\mathrm{i}} = (x_{i1}, x_{i2}, ...x_{ip})^T$ are the input variables and $y_i$ are the output variables. Let $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1 \hat{\beta}_2 ... \hat{\beta}_p\right)^T$, the Lasso estimate $(\hat{\alpha}, \hat{\beta})$ is defined by:

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin} \sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \quad (7)$$

$$\text{subject to } \sum_{\mathrm{j}} \left|\beta_{\mathrm{j}}\right| \le t \quad (8)$$

where $t \ge 0$ is a tuning parameter, which controls the amount of shrinkage applied to the estimates. Equation (8) is equivalent to minimizing:

$$\frac{1}{2n} \sum (y_i - \alpha - \beta x_{\mathrm{i}})^2 + \lambda \sum \left|\beta_j\right| \quad (9)$$

which is a least squares criterion with a penalty specified by $\lambda$ for large coefficient estimates. $\sum \left|\beta_j\right|$ is the coefficient vector, which delivers a spares solution vector $\beta_\lambda$; a larger $\lambda$ results that more elements of $\beta_\lambda$ are zero. If $\lambda = 0$, the Lasso is the same as ordinary least squares; as $\lambda$ increases, shorter vectors are preferred.

### 4.2. Stock Prediction Module

SVM was first presented by Vapnik (1995) and it aims to learn a maximum margin hyperplane to divide training instances into disjoint groups (Vapnik, 1995). Because of advantages in solving the classification, regression and time series problems, especially in the case of small sample, nonlinearity and high-dimension, SVM has been used in various studies (Hu et al., 2013; Huang, 2012; Lee, 2009). Because stock market data are highly noisy and complex in dimension (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998; Huang, 2012), SVM is promising to obtain good predictive performance in this study.

Given a dataset $S$ of $n$ labeled training instances $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$, $x_i \in R^k$, $i=1,...,n, y_i \in \{0,1\}$, where $x_i$ is the input vector, $y_i$ is the output variables and $k$ means the input dimension. Each training instance $x_i$ belongs to either of the two classes according to its label $y_i$.

In the non-linear case, the maximum margin hyperplane can be defined by the separate function as follows:

$$y = b + \sum w_i y_i K(x(i), x) \tag{10}$$

where $x$ means a test instance, $x(i)$ represents the support vectors, and $b$ and $w_i$ are the parameters learned by SVM. $K(x(i),x)$ is the kernel function that generates inner products and it enables SVM to support different types of nonlinear decision surfaces in the input space. There are two common kernel functions - the polynomial kernel and the Gaussian radial basis function, which respectively defined as:

$$K(x,y) = (xy+1)^d \tag{11}$$

and:

$$K(x,y) = \exp(-1/\delta^2(x-y)^2) \tag{12}$$

where $d$ is the degree of the polynomial kernel and $\delta^2$ represents the bandwidth of the Gaussian radial basis function.

To obtain the maximum separation of the two classes of instances, SVM find an optimal hyperplane by solving the convex quadratic programming problem:

$$\text{Min } \frac{1}{2}\|w\|^2 \tag{13}$$

$$\text{subject to } y_i(w \cdot x_i + b) \geq 1, i = 1,\ldots,n \tag{14}$$

For the construction of stock prediction module, a set of popular financial indicators are selected as input variables. These indicators are listed in Table 1 and all of them are derived from either academic papers or practical experience.

In the process of model training, each training instance is assigned with a class label that indicates whether having a positive excess return. A positive (negative) excess return means that the difference between the stock return and the benchmark return is positive (negative), or that a stock outperforms (underperforms) the benchmark. Although some studies classify instances into three states, for example, uptrend, downtrend and steady, by setting a threshold (Chang et al., 2011; Lai et al., 2009), no consensus on the most proper threshold exists and trials on different thresholds suffer from data-snooping bias. Therefore, this study adopts the most straightforward way (i.e., positive/negative excess return).

In the process of model application, the system generates a buy (sell) signal whenever the stock prediction model predicts a stock would outperform (underperform) the benchmark. Subsequently, these signals are processed by other modules to generate the final trading decision.

### 4.3. Portfolio Optimization Module

If neither the required rate of return nor the risk tolerance is specified, the solution of Markowitz' mean-variance model can be transformed to the problem of optimizing the risk-adjusted expected return. The required expected return can be replaced by expected return minus the square root of risk:

$$r_p - A\sigma_p \tag{15}$$

where multiplier $A$ is a risk aversion parameter specified by users. A larger $A$ implies that an

*Table 1. Hybrid strategy model factors*

| Category | Factors | Description | Ref. |
|---|---|---|---|
| Price rationality | (1) EP ratio | Earnings-to-price ratio | (Mukherji, Dhatt, & Kim, 1997) |
| | (2) BP ratio | Book-to-price ratio | (Mukherji et al., 1997) |
| | (3) SP ratio | Sales-to-price ratio | (Mukherji et al., 1997) |
| Profitability | (4) ROE | Return on equity (after tax) | (Omran, 2004) |
| | (5) ROA | Return on asset (after tax) | (Omran, 2004) |
| Growth | (6) OIG | Operating income growth | (Ikenberry & Lakonishok, 1993) |
| | (7) NIG | Net income growth | (Sadka & Sadka, 2009) |
| | (8) ONCF | Operating net cash flow year-on-year growth rate | |
| Leverage | (9) DE ratio | Debt-to-equity ratio | (Omran, 2004) |
| | (10) ALR | Asset liability ratio | |
| Liquidity | (11) CR | Current ratios | (Omran, 2004) |
| Operating | (12) FATR | Fixed assets turnover rate | |
| | (13) TATR | Total assets turnover rate | |
| | (14) ITR | Interest turnover rate | (Omran, 2004) |
| | (15) TAT | Total asset turnover year-on-year | |
| Scale | (16) Ln(S) | The lagged market value | (Fama & French, 1992) |
| | (17) Ln(M) | The lagged circulation market value | |
| | (18) Ln(A) | The lagged asset | (Fama & French, 1992) |
| Cash flow | (19) NCFOC/TMV | Net cash flows of operating activities-to-total market value | |

investor is more sensitive to risk. For example, if $A = 2$, the investor considers that 1% rise in risk (i.e., variance of return) equals to 2% decline in expected return. In this study, $A = 2$ represents an aggressive investment style, while $A = 4$ is a defensive investment style. An aggressive investor will pursue extra expected return at the cost of higher risk.

Given formula (15), the previous mean-variance problem can be expressed as a risk-adjusted portfolio optimization problem, as follows:

$$\text{Max} \quad r_p - A\sigma_p = \sum_{i=1}^{n} w_i r_i - A\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij} \tag{16}$$

$$\text{subject to} \quad \sum_{i=1}^{n} w_i = 1 \tag{17}$$

## 4.4. Other Portfolio Management Modules

Other than the portfolio optimization module, remaining modules such as the risk monitoring module, the portfolio tracking module and the benchmark analysis module are also important to portfolio management.

For portfolio performance and risk evaluation, three popular measures were selected, including return ratio, annualized return and Sharpe ratio. As a risk indicator, Sharpe ratio

measures the average excess return on a risk-free return per unit of standard deviation in an investment asset or a trading strategy (Sharpe, 1994). A higher Sharpe ratio implies higher return and lower volatility, which is preferred.

For benchmark analysis, besides the market benchmark index, iTrade has incorporated strategies of three world-wide master investors, including Warren E. Buffett, William J. O'Neil and Richard Driehaus. Their linguistic theories can be quantified following the approach proposed in (Huang, 2009; Huang & Jane, 2009), and modified in accordance with the selected data source. The quantified strategies of Buffett (Hu et al., 2013), O'Neil and Driehaus are shown in Table 2, Table 3 and Table 4, respectively. Introductions of these investment philosophies and similar quantified investment rules can be found at the website *FINASIA* (http://www.tej.com.tw/twsite/Default.aspx?TabId=389).

## 5. APPLICATION CASE OF ITRADE

This section provides an application case to demonstrate and evaluate iTrade. A back-testing was conducted under a simulation environment for stock trading in the Chinese Stock Exchanges.

### 5.1. System Setting

Stocks of two industries (i.e., nonferrous metal industry and commerce and trade industry) are randomly selected for back-testing. Details of these two industries are shown in Table 5. Annual fundamental data and price observations of

*Table 2. Quantified Buffett investment strategy*

| # | Factor |
|---|---|
| 1 | Company's gross profit margin > Industry average (gross profit margin) |
| 2 | Five-year average of shareholders' equity ratio > 15% |
| 3 | Shareholders' equity ratio > Industry and market average (shareholders' equity ratio) |
| 4 | Debt ratio < stock pool average (debt ratio) |
| 5 | Price / (free cash flow per share) < 10 |
| 6 | Cash flow per share >= earnings per share |

*Table 3. Quantified O'Neil investment strategy*

| # | Factor |
|---|---|
| 1 | The year-on-year growth rate of EOCPS on last fiscal quarter >= 20% |
| 2 | The year-on-year growth rate of EPS on last fiscal quarter > 0 |
| 3 | The year-on-year growth rate of EOCPS on last fiscal quarter > The year-on-year growth rate of EOCPS on last but one fiscal quarter |
| 4 | EOCPS on last two fiscal quarter > 0 |
| 5 | EOCPS on last year >= EPS on last year |
| 6 | The growth rate of EOCPS on last three year >= 25% |
| 7 | The growth rate of EOCPS on every last three year > 0 |
| 8 | The EPS of last year ranking in the top 30% among all the stocks |

-EOCPS: Earnings of operating activities per share.
-EPS: Earnings per share.

*Table 4. Quantified Driehaus investment strategy*

| # | Factor |
|---|---|
| 1 | Total asset<the average of market |
| 2 | The expected growth of operating income > the value of the same period in last year |
| 3 | The expected growth of earnings after tax > the value of the same period in last year |
| 4 | The growth of operating income on recent fiscal quarter > the value of the same period in last year |
| 5 | The growth of earnings after tax on recent fiscal quarter > the value of the same period in last year |
| 6 | The turnover of volume on recent 30 trading days > the average of market |
| 7 | The change rate of relative strength of stock price on recent 30 trading days > 0 |

*Table 5. List of the selected industries*

| # | Industry | Num. of Stocks |
|---|---|---|
| 1 | Nonferrous metal | 63 |
| 2 | Commerce and Trade | 81 |

these stocks were fetched from the well-known Chinese financial data source, *Guotaian Finance Database*, covering a twelve-year period from 2000 to 2011. Besides, Shanghai composite Index is selected as the market benchmark for the calculation of excess returns, which is in accordance with the industry practice.

The iTrade uses sliding window method to divide the sample data into different groups of training and test data, and the length of training data to the length of testing data was set to 5:1 in this application case. For instance, training data of the first group are from the year of 2000 to 2004 and testing data of this group are from the succeeding year of 2005. And training data of the last group are from 2006 to 2010, and the corresponding testing data are from 2011. Therefore, the model is trained and tested for seven different time frames.

As for risk-adjusted portfolio optimization, two levels of risk aversion parameter were examined. Risk aversion parameter = 2 indicates an aggressive investment style, while risk aversion parameter = 4 represents a defensive investment style. Dynamic weights in a portfolio are calculated based on historical weekly return data.

## 5.2. Results

Figure 4 and Figure 5 show the seven-year back-testing cumulative returns, under different risk aversion scenarios, on the nonferrous metal industry and the commerce and trade industry, respectively. Table 6 reports performance details of different risk aversion scenarios, compared with those of three world-famous master investors. The non-risk-adjusted case constructed portfolio with equal weights, while the risk-adjusted case constructed portfolio with dynamic weights that were calculated by the portfolio optimization module according to the specified risk aversion parameter ($A$).

As Table 6 shows, iTrade was able to identify up-trend stocks and generate considerable excess return (return ratio are 4.91 and 8.63 in the nonferrous metal industry and the commerce and trade industry, respectively) with a significant higher return ratio than that of the market benchmark (2.23, buy-and-hold strategy on Shanghai Composite Index).

In the non-risk-adjusted case, iTrade surpassed both Buffett's and O'Neil's strategies in terms of return ratio and close/higher Sharpe ratio. However, Driehaus' strategy performed

*Figure 4. Comparison of cumulative returns (nonferrous metal industry)*

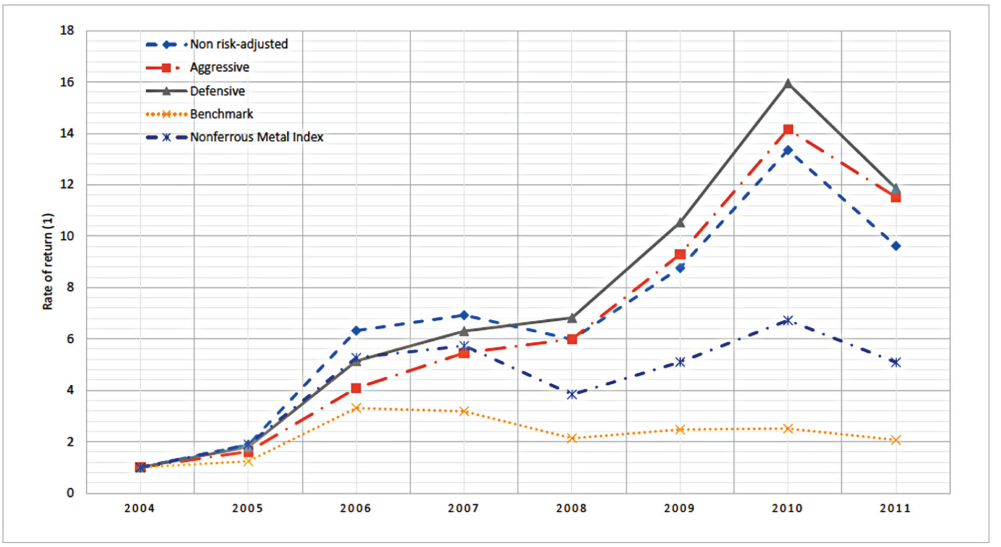
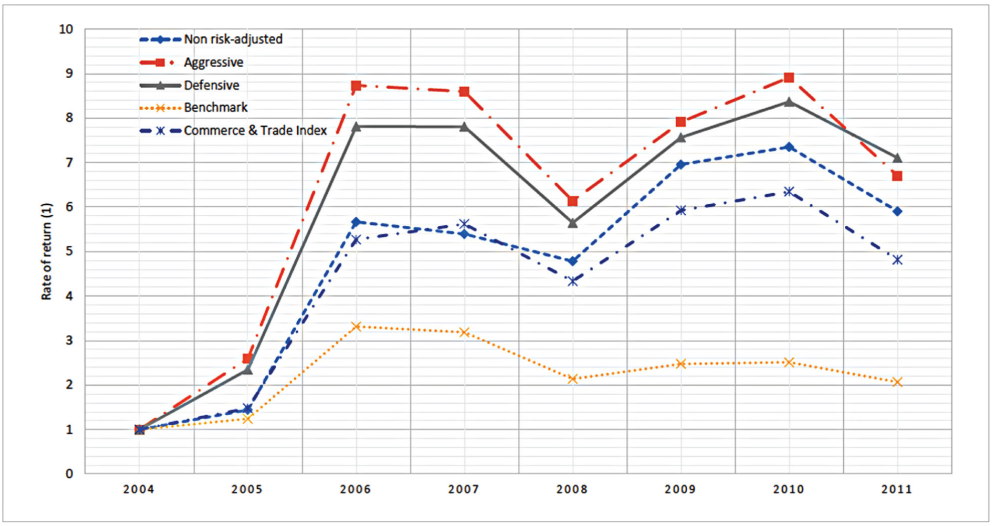



*Figure 5. Comparison of cumulative returns (commerce and trade industry)*




much better in terms of both return ratio (12.16) and Sharpe ratio (0.68).

Notably, both two risk-adjusted cases performed better than the non-risk-adjusted case, which implies that risk-adjusted portfolio optimization can significantly improve portfolio performance with higher return ratio and Sharpe ratio.

Considering practical requirement of different investment styles, some investors are willing to suffer higher risk level for the sake of higher return, while others prefer much safer

*Table 6. Performance of different strategies*

| | Method | Stock Pool | Return | Annualized Return | Sharpe Ratio |
|---|---|---|---|---|---|
| Non-risk-adjusted | iTrade | Nonferrous metal | 4.91 | 0.29 | 0.46 |
| | | Commerce and trade | 8.63 | 0.38 | 0.64 |
| | Buffett | A-Shares | 3.47 | 0.12 | 0.48 |
| | O'Neil | A-Shares | 4.57 | 0.18 | 0.62 |
| | Driehaus | A-Shares | 12.16 | 0.82 | 0.68 |
| Risk-adjusted ($A = 2$) | iTrade | Nonferrous metal | 5.71 | 0.31 | 0.54 |
| | | Commerce and trade | 10.52 | 0.42 | 0.94 |
| Risk-adjusted ($A = 4$) | iTrade | Nonferrous metal | 6.11 | 0.32 | 0.56 |
| | | Commerce and trade | 10.87 | 0.42 | 0.81 |
| Market Benchmark | Buy-and-hold | | 2.23 | 0.12 | - |

portfolios even at the cost of lower returns. Results of two risk-adjusted scenarios showed that iTrade could provide investors such a choice by balancing between return and risk, through the risk aversion parameter. As shown in Table 6, risk aversion parameters had different effects on performance. Risk aversion parameter $A = 4$ (defensive investment style) produced slightly higher returns (6.11 and 10.87) compared with risk aversion parameter $A = 2$ (aggressive investment style) (with returns of 5.71 and 10.52). This result is surprising but reasonable. Many people may have a sense that aggressive investment style could bring higher return because of the compensation of the higher risk. However, our experimental results showed that defensive investment style defeated the aggressive one. This conclusion may not be widely applicable because selection bias exist in back-testing where it only included the nonferrous metal industry and the commerce and trade industry of China Stock Exchanges during the year of 2000 to 2011.

## 5.3. Concept Drift Analysis

Concept drift analysis may be the most valuable work in stock market analysis. One rule in market cannot always make money. Stock market conditions are ever-changing, thus trading rules must be timely adapted to the current situation. This would incur higher profit and lower risk.

According to Table 7, existence of concept drift between different years and different industries was obvious. Several factors had good explanatory power in some years but less in the other years.

As for the *nonferrous* metal industry, factor sets were stable. Just a few factors (#1, 2, 3, 9, 16, 17 and 19) were steady, while the others varied irregularly.

As for the commerce and trade industry, factor sets appeared to be more complex and instable. This implied that investors of this industry should pay more attention to market information that may have significant impact on the stock price.

Moreover, obvious difference was found between different stock market conditions, e.g., between bear and bull markets. Consider the nonferrous metal industry for example. In the bull market year of 2006, only two price rationality factors (*E/P ratio* and *S/P ratio*) were found informative, while in the bear market year of 2008, besides these two factors, another price rationality factor (*B/P ratio*), a leverage factor

*Table 7. Comparison of feature subsets over seven-year testing period*

| Category | Factors | Nonferrous Metal Industry | | | | | | | Commerce and Trade Industry | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 05 | 06 | 07 | 08 | 09 | 10 | 11 | 05 | 06 | 07 | 08 | 09 | 10 | 11 |
| Price rationality | (1) E/P | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● |
| | (2) B/P | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| | (3) S/P | ● | ● | ● | ● | ● | ● | ● | | ● | ● | ● | ● | ● | ● |
| Profitability | (4) ROE | ● | | | | ● | ● | ● | | ● | | | ● | ● | ● |
| | (5) ROA | ● | | | | ● | | | | | | | ● | | ● |
| Growth | (6) OIG | ● | | | | | | | | | | | | | |
| | (7) NIG | ● | | | | | ● | ● | | ● | | | ● | ● | |
| | (8) ONCF | | | | | | | | | ● | | | ● | ● | |
| Leverage | (9) DE | | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| | (10)ALR | | | | | | | | ● | ● | ● | ● | ● | ● | ● |
| Liquidity | (11) CR | | | | | | | ● | ● | ● | | | ● | ● | |
| Operating | (12) FATR | ● | | | | | | ● | | ● | | | ● | ● | |
| | (13) TATR | | | | | | | ● | ● | ● | ● | | ● | ● | |
| | (14) ITR | | | | | | | ● | | ● | | | ● | ● | ● |
| | (15) TAT | | | | | | | ● | | | | | | | |
| Scale | (16) Ln(S) | ● | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| | (17) Ln(M) | | | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| | (18) Ln(A) | | | | | | | | | ● | ● | ● | ● | ● | |
| Cash flow | (19) NCFOC/TMV | | | ● | ● | ● | | ● | | ● | | | ● | ● | ● |

(*Debt-to-equity ratio*) and two scale factors (*Ln(S)* and *Ln(M)*) must be added into consideration. Consider another example based on the commerce and trade industry, which excludes manufacturing corporates. Much more factors were selected in the year of 2006 compared to 2008, which is a sharp contrast to the case of nonferrous metal industry.

## 6. CONCLUSION AND FUTURE WORK

Stock market data are complex, non-linear and time-variant. Individual investors, especially the non-professional investors, are in great need of a mobile intelligent trading decision support system. Compared to traditional standalone stock analysis software, the proposed iTrade can satisfy this need as it runs on a C/S system architecture and has incorporated various technologies of data mining, concept drift adaptation and portfolio optimization. Specifically, iTrade is characterized by 1) a data-driven intelligent learning model, 2) a concept drift adaptation process, and 3) a rigorous benchmark analysis. In addition, iTrade enables user access through a variety of mobile devices, e.g. smartphones tablets and notebooks, and this is especially important to non-professional investors that are on a business trip or in an offline working place.

Application case showed that iTrade performed well in terms of both cumulative return ratio and Sharpe ratio, compared to the Buy-and-Hold strategy on the Shanghai Composite Index and the quantified strategies of world-famous master investors. As a highlight of this study,

the concept drift analysis revealed interesting findings on the time-variant characteristics of the most informative financial factors for stock prediction on the nonferrous metal industry and the commerce and trade industry.

Directions of future research include 1) conducting more comprehensive concept drift analysis for further insight, such as (Hu, Feng, Zhang, Ngai, & Liu, 2014), and incorporating more advanced feature selection/causal analysis methods, such as (Hu, Zhang, Ngai, Cai, & Liu, 2013; Zhang et al., 2014), 2) carrying out more rigorous comparative studies with strategies of other world-famous master investors or fund managers, and 3) examining iTrade in other stock markets.

## ACKNOWLEDGMENT

This research was partly supported by the National Natural Science Foundation of China (71271061, 70801020), Science and Technology Planning Project of Guangdong Province, China (2010B010600034, 2012B091100192), and The Ministry of Education Innovation Team Development Plan, Guangdong Natural Science Foundation Research Team, and Business Intelligence Key Team of Guangdong University of Foreign Studies (S2013030015737, IRT1224, TD1202).

## REFERENCES

Armano, G., Murru, A., & Roli, F. (2002). Stock market prediction by a mixture of genetic-neural experts. *International Journal of Pattern Recognition and Artificial Intelligence*, *16*(5), 501–526. doi:10.1142/S0218001402001861

Ashrafi, M. Z., Taniar, D., & Smith, K. (2007). Redundant association rules reduction techniques. *International Journal of Business Intelligence and Data Mining*, *2*(1), 29–63. doi:10.1504/IJBIDM.2007.012945

Chang, P., Fan, C., & Lin, J. (2011). Trend discovery in financial time series data using a case based fuzzy decision tree. *Expert Systems with Applications*, *38*(5), 6070–6080. doi:10.1016/j.eswa.2010.11.006

Chen, Y., & Cheng, C. (2009). Evaluating industry performance using extracted rgr rules based on feature selection and rough sets classifier. *Expert Systems with Applications*, *36*(5), 9448–9456. doi:10.1016/j.eswa.2008.12.036

Chen, Y., Ohkawa, E., Mabu, S., Shimada, K., & Hirasawa, K. (2009). A portfolio optimization model using genetic network programming with control nodes. *Expert Systems with Applications*, *36*(7), 10735–10745. doi:10.1016/j.eswa.2009.02.049

Chun, S., & Park, Y. (2005). Dynamic adaptive ensemble case-based reasoning: Application to stock market prediction. *Expert Systems with Applications*, *28*(3), 435–443. doi:10.1016/j.eswa.2004.12.004

Daly, O., & Taniar, D. (2004). Exception rules mining based on negative association rules. In LaganáA. GavrilovaM.KumarV.MunY.TanC. J. K.GervasiO. (Eds.), *Proceedings of the International Conference on Computational Science and Its Applications (ICCSA 2004), Part IV, Lecture Notes in Computer Science* (Vol. 3046, pp. 543-552). Berlin/Heidelberg: Springer. doi:10.1007/978-3-540-24768-5_58

Delany, S. J., Cunningham, P., Tsymbal, A., & Coyle, L. (2005). A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems*, *18*(4–5), 187–195. doi:10.1016/j.knosys.2004.10.002

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*(4), 927–940. doi:10.1016/j.eswa.2005.06.024

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427–465. doi:10.1111/j.1540-6261.1992.tb04398.x

Goh, J., & Taniar, D. (2004). Mining frequency pattern from mobile users. In M. Negoita, R. Howlett, & L. Jain (Eds.), Lecture Notes in Computer Science: Vol. 3215. *Knowledge-Based Intelligent Information and Engineering Systems (8th Knowledge-Based Intelligent Information and Engineering Systems, KES 2004, Part III)* (pp. 795–801). Berlin, Heidelberg: Springer.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *Intelligent Systems and their Applications, IEEE, 13*(4), 18-28

Hsu, N., Hung, H., & Chang, Y. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis*, *52*(7), 3645–3657. doi:10.1016/j.csda.2007.12.004

Hu, Y., Feng, B., Zhang, X., Ngai, E. W. T., & Liu, M. (2014). (in press). Stock trading rule discovery with an evolutionary trend following model. *Expert Systems with Applications*. doi:10.1016/j.eswa.2014.07.059

Hu, Y., Feng, B., Zhang, X., Qiu, X., Li, R., & Xie, K. (2013). *A prediction model for stock market: a comparison of the world's top investors with data mining method*. Paper presented at Twelfth Wuhan International Conference on E-business (WHICEB2013), Wuhan, China.

Hu, Y., Zhang, X., Ngai, E. W. T., Cai, R., & Liu, M. (2013). Software project risk analysis using bayesian networks with causality constraints. *Decision Support Systems*, *56*, 439–449. doi:10.1016/j.dss.2012.11.001

Huang, C. (2012). A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, *12*(2), 807–818. doi:10.1016/j.asoc.2011.10.009

Huang, K. Y. (2009). Application of vprs model with enhanced threshold parameter selection mechanism to automatic stock market forecasting and portfolio selection. *Expert Systems with Applications*, *36*(9), 11652–11661. doi:10.1016/j.eswa.2009.03.028

Huang, K. Y., & Jane, C. (2009). A hybrid model for stock market forecasting and portfolio selection based on arx, grey system and rs theories. *Expert Systems with Applications*, *36*(3, Part 1), 5387–5392. doi:10.1016/j.eswa.2008.06.103

Ikenberry, D., & Lakonishok, J. (1993). Corporate governance through the proxy contest: Evidence and implications. *The Journal of Business*, *66*(3), 405–435. doi:10.1086/296610

Kuncheva, L. (2008). *Classifier ensembles for detecting concept change in streaming data: overview and perspectives*. Paper presented at the 2nd Workshop SUEMA 2008 (ECAI 2008), Patras, Greece.

Lai, R. K., Fan, C., Huang, W., & Chang, P. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications*, *36*(2, Part 2), 3761–3773. doi:10.1016/j.eswa.2008.02.025

Lee, M. (2009). Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, *36*(8), 10896–10904. doi:10.1016/j.eswa.2009.02.038

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*(1), 77–91.

Mukherji, S., Dhatt, M. S., & Kim, Y. H. (1997). A fundamental analysis of korean stock returns. *Financial Analysts Journal*, *53*(3), 75–80. doi:10.2469/faj.v53.n3.2086

Na, S. H., & Sohn, S. Y. (2011). Forecasting changes in korea composite stock price index (kospi) using association rules. *Expert Systems with Applications*, *38*(7), 9046–9049. doi:10.1016/j.eswa.2011.01.025

O, J., Lee, J., Lee, J. W., & Zhang, B. (2006). Adaptive stock trading with dynamic asset allocation using reinforcement learning. *Information Sciences, 176*(15), 2121-2147

Omran, M., & Ragab, A. (2004). Linear versus nonlinear relationships between financial ratios and stock returns: Empirical evidence from egyptian firms. *Review of Accounting and Finance*, *3*(2), 84–102. doi:10.1108/eb043404

Ren, N., Zargham, M., & Rahimi, S. (2006). A decision tree-based classification approach to rule extraction for security analysis. *International Journal of Information Technology & Decision Making*, *5*(1), 227–240. doi:10.1142/S0219622006001824

Sadka, G., & Sadka, R. (2009). Predictability and the earnings–returns relation. *Journal of Financial Economics*, *94*(1), 87–106. doi:10.1016/j.jfineco.2008.10.005

Sharpe, W. F. (1994). The sharpe ratio. *Journal of Portfolio Management*, *21*(1), 49–58. doi:10.3905/jpm.1994.409501

Taniar, D., Rahayu, W., Lee, V., & Daly, O. (2008). Exception rules in association rule mining. *Applied Mathematics and Computation*, *205*(2), 735–750. doi:10.1016/j.amc.2008.05.020

Thawornwong, S., & Enke, D. (2004). The adaptive selection of financial and economic variables for use with artificial neural networks. *Neurocomputing*, *56*, 205–232. doi:10.1016/j.neucom.2003.05.001

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *58*(1), 267–288.

Tsai, C., & Hsiao, Y. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. *Decision Support Systems*, *50*(1), 258–269. doi:10.1016/j.dss.2010.08.028

Tsai, C., Lin, Y., Yen, D. C., & Chen, Y. (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing*, *11*(2), 2452–2459. doi:10.1016/j.asoc.2010.10.001

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York: Springer-Verlag. doi:10.1007/978-1-4757-2440-0

Wang, F., & Xu, Y. (2004). What determines chinese stock returns? *Financial Analysts Journal*, *60*(6), 65–77. doi:10.2469/faj.v60.n6.2674

Wang, H., Fan, W., Yu, P. S., & Han, J. (2003). *Mining concept-drifting data streams using ensemble classifiers*. Paper presented at the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi:10.1145/956755.956778

Wang, H., Li, G., & Tsai, C. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *69*(1), 63–78. doi:10.1111/j.1467-9868.2007.00577.x

Wang, Z. T., & Tan, S. H. (2009). Identifying idiosyncratic stock return indicators from large financial factor set via least angle regression. *Expert Systems with Applications*, *36*(4), 8350–8355. doi:10.1016/j.eswa.2008.10.018

Zhang, X., Hu, Y., Xie, K., Wang, S., Ngai, E. W. T., & Liu, M. (2014). A causal feature selection algorithm for stock prediction modeling. *Neurocomputing*, *142*, 48–59. doi:10.1016/j.neucom.2014.01.057

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zliobaite, I. (2009). Learning under concept drift: an overview: Faculty of Mathematics and Informatics, Vilnius University.

*Yong Hu is currently a Professor and Director of Institute of Business Intelligence and Knowledge Discovery at the Guangdong University of Foreign Studies and Sun Yat-Sen University. He received his B.Sc in Computer Science, M.Phil and Ph.D. in Management Information Systems from Sun Yat-Sen University. His research interests are in the areas of business intelligence, quantitative investment, medical informatics, software project risk management, Spam filtering and decision support systems. He has published more than 50 papers in journals and conferences such as DSS, JASIST, IJPR, ESWA, IST, IEEE ICDM, and others. Dr. Hu's research is supported by the National Science Foundation of China, the Science and Technology Planning Project of Guangdong Province, and the Key Team of Business Intelligence from Guangdong University of Foreign Studies.*

*Xiangzhou Zhang is a Ph.D. student in Sun Yat-sen University and working as an assistant researcher in Institute of Business Intelligence and Knowledge Discovery at the Guangdong University of Foreign Studies and Sun Yat-Sen University. He has received his BSc in Computer Science from Sun Yat-Sen University, M.Phil in Guangdong University of Foreign Studies. His research interests are quantitative investment, design science, software project risk management and business intelligence. He has published papers in a number of international journals and conferences including* Decision Support Systems, Knowledge-based Systems, Expert Systems with Applications, Neurocomputing, International Journal of Data Warehousing and Mining, Journal of Software, FSKD'09, EIDWT2013 and WHICEB2013.

*Bin Feng is an M.Phil student in Guangdong University of Foreign Studies and working as an assistant researcher in Institute of Business Intelligence and Knowledge Discovery at the Guangdong University of Foreign Studies. He has received his BSc in Computer Science from Zhengzhou University. His research interests are quantitative investment and business intelligence. He has published papers in international journals and conferences including* Expert Systems with Applications, International Journal of Data Warehousing and Mining, EIDWT2013 and WHICEB2013.

*Kang Xie is currently a professor of management science, School of Business, Sun Yat-sen University. He is Standing Associate Director-General of China Information Economics Society, Standing Director of China Association for Information Systems, and Advanced Consultant of Ministry of Commence on Electronic Commence. He received his Ph.D. in management from Renmin University of China. He has published more than 11 books and numerous papers in a number of international journals and conferences. His research interests are in the areas of management science, Information Economics, E-commerce Economics, and Enterprise informatization.*

*Mei Liu is Assistant Professor of Medical Informatics in the Department of Internal Medicine at University of Kansas Medical Center. She currently also serves as an Associate Editor of Decision Support Systems. Dr. Liu received her Ph.D. degree in Computer Science from the University of Kansas and completed her postdoctoral training in medical informatics in the Department of Biomedical Informatics at Vanderbilt University. Her research interest includes healthcare risk management, medical informatics, data mining, and machine learning. She has published numerous articles in top journals such as* JAMIA, Bioinformatics, Decision Support Systems and PLoS ONE.