MACHINE-LEARNING CLASSIFICATION TECHNIQUES FOR THE ANALYSIS AND PREDICTION OF HIGH-FREQUENCY STOCK DIRECTION

by

Michael David Rechenthin

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Business Administration (Management Sciences)
in the Graduate College of
The University of Iowa

May 2014

Thesis Supervisor: Professor W. Nick Street

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Michael David Rechenthin

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Business Administration (Management Sciences) at the May 2014 graduation.

Thesis committee:

W. Nick Street, Thesis Supervisor

Gautam Pant

Padmini Srinivasan

Tong Yao

Kang Zhao

ACKNOWLEDGEMENTS

I would like to dedicate this thesis to my wife, Abby, and to my parents, Dr. and Mrs. David and Ellen Rechenthin. I am deeply grateful for my dad's valued guidance throughout this entire process and for the loving support and encouragement from Abby and my mom.

I would like to express my gratitude for the supervision of my advisor, Dr. Nick Street, whose direction made this paper possible. I would also like to thank my thesis committee, Dr. Padmini Srinivasan, Dr. Gautam Pant, Dr. Kang Zhao, and Dr. Tong Yao, for their insightful comments and advice.

Finally, I would like to acknowledge the generous financial support of University of Iowa's Department of Management Sciences.

ABSTRACT

This thesis explores predictability in the market and then designs a decision support framework that can be used by traders to provide suggested indications of future stock price direction along with an associated probability of making that move. Markets do not remain stable and approaches that are highly predictive at one moment may cease to be so as more traders spot the patterns and adjust their trading techniques. Ideally, if these "concept drifts" could be anticipated, then the trader could store models to use with each specific market condition (or concept) and later apply those models to incoming data. The assumption however is that the future is uncertain, therefore future concepts are still undecided.

Maintaining a model with only the most up-to-date price data is not necessarily the most ideal choice since the market may stabilize and old knowledge may become useful again. Additionally, decreasing training times to enable modified classifiers to work with streaming high-frequency stock data may result in decreases in performance (e.g. accuracy or AUC) due to insufficient learning times. Our framework takes a different approach to learning with drifting concepts, which is to assume that concept drift occurs and builds this into the model. The framework adapts to these market changes by building thousands of traditional base classifiers (SVMs, Decision Trees, and Neural Networks), using random subsets of past data, and covering similar (sector) stocks and heuristically combining the best of these base classifiers. This "ensemble", or pool of multiple models selected to achieve better predictive per-

formance, is then used to predict future market direction. As the market moves, the base classifiers in the ensemble adapt to stay relevant and keep a high level of model performance. Our approach outperforms existing published algorithms.

This thesis also addresses problems specific to learning with stock data streams, specifically class imbalance, attribute creation (e.g. technical and sentiment analysis), dimensionality reduction, and model performance due to release of news and time of day. Popular methods for dealing with each are discussed.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Predicting stock price direction is something individuals and financial firms have researched for years. Books and papers have been written on the subject, but rarely are the results repeatable. Recent research has shown greater predictability in high-frequency stock data (by second or by minute, rather than daily or weekly), but this research is often under represented in the academic literature. Historically, this is due to the lack of availability of trade-by-trade data, and the difficulty in working with such large quantities of it. Furthermore, determining future market direction in practice requires special consideration since streaming stock data may arrive faster than a model may produce results; a model that takes 30 minutes to arrive at a prediction is of little value if the objective was to predict one minute in the future. This thesis explores the predictability of stock market direction using machine learning classification techniques and high-frequency stock data.

These techniques would be of considerable interest to quantitative traders who produce mathematical models that account for as much of 55% of the total volume of US traded stocks [2]. Our research objective is to build a decision support framework that can be used by traders to provide suggested indications of future stock price direction along with an associated probability of making that move. For example, if a trader wanted to purchase an equity position, knowing whether to buy now or wait and re-evaluate in n seconds could allow the trader to purchase the stock at a lower price than was previously expected. Over time this could make a significant

difference in the profitability of the strategy.

It is argued that the lack of published working models exists because there is little incentive to publish such methods in academic literature. The incentive to instead sell them to a trading firm is much greater, monetarily. Also Timmermann and Granger [216] write of a possible "file drawer" bias in published studies due to the difficulty in publishing empirical results that are often barely or border-line statistically insignificant; but markets, since they are partially driven by human emotions, involve a large degree of error. This may result in a glut of research arguing that the market is efficient, and thus unpredictable; Timmermann and Granger calls this the "reverse file-drawer" bias. It is also possible that many traditional forms of stock market prediction are simply inadequate or sponsoring companies may not wish to divulge successful applications [245]. We demonstrate that inefficiency and moments of predictability exist using 22 million stock transactions. This is discussed further in Chapter 2.

When predicting stock price direction, practitioners typically use one of three approaches. The first is the fundamental approach, which examines the economic factors that drive the price of stock (e.g. a company's financial statements such as the balance sheet or income statement). The second approach is to use traditional technical analysis to anticipate what *others* are thinking based on the price and volume of the stock. Indicators are computed from past prices and volumes and these are used to foresee future changes in prices. The goal of technical analysis is to identify regularities by extracting patterns from noisy data and by inspecting the stock charts

visually. Studies show 80% to 90% of polled professionals and individual investors rely on at least some form of technical analysis [157, 162, 163, 213]. With recent breakthroughs in technology and algorithms, technical analysis has morphed into a more quantitative and statistical approach [154]; this is what we call quantitative technical analysis and it is the third approach to predicting market direction. Whereas traditional technical analysis is *visual*, quantitative technical analysis is *numerical*, which allows us to easily program the rules into a computer. This is the method explored in this thesis.

Markets do not remain stable; indicators that be highly predictive at one moment may cease to be so as more traders spot the patterns and implement them in their trading approaches. Widespread adoption of a particular trading strategy is enough to drive the price either up or down enough to eliminate the pattern [201, 215]. This *concept drift* complicates the learning of models and is unique to streaming data. As the concept changes, model performance may decrease, requiring an update in the training data and/or change in the quantitative technical analysis indicators used as attributes. Modern machine learning classification techniques provide solutions and this, along with quantitative technical analysis, allows us to outperform existing published methods of stock market direction.

We begin Chapter 3 with an introduction to the discussion of machine learning for streaming data and in particular high-frequency stock data. This includes descriptions of traditional supervised learning methods and different ways of evaluating classifiers that are used for analysis later in the thesis. Chapter 4 then discusses

the two main approaches to learning from streaming data: adaptive (online) and wrapper methods. Adaptive methods learn data incrementally (as instances arrive) and efficiently with a single pass through the data; they are ideal for use with high-frequency streaming data because they require limited time and memory. *Forgetting factors* can be integrated in the model to give less weight to older data, thus *gradually* making older data obsolete. Wrapper methods use traditional classification algorithms, such as support vector machines or neural networks (both are explained in Chapter 3), that learn on collected batches of data. The models are then chosen and combined to form predictions, such as through the use of ensembles.

High-frequency streaming stock data requires special consideration for three main reasons: first, the market is constantly changing, so models quickly become obsolete; second, speedy processing is required to make fast predictions; and third, specific volumes of data are needed to make precise decisions. In Chapter 5 we address problems specific to stock data such as imbalanced datasets, too much and irregularly spaced data, and attribute creation and selection.

In Chapter 6, we discuss our new wrapper-based ensemble method that provides a solution for all three considerations outlined in Chapter 5. This method is created by building thousands of models in parallel using price and volume data covering different periods of time and using different stocks, and then efficiently switching between these models as time progresses. We then demonstrate an improvement over existing methods using our new approach.

Lastly in Chapter 7 we summarize our thesis and discuss an additional idea

for future work; specifically using misclassification costs to optimize decisions rather than AUC.