

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

by

Michael David Rechenhain

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

UMI Number: 3628434

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3628434

Published by ProQuest LLC (2014). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against unauthorized copying under Title 17, United States Code



ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Copyright by
MICHAEL DAVID RECHENTHIN
2014
All Rights Reserved

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 Rechenhain

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.

PREVIEW

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	xii
CHAPTER	
1 INTRODUCTION	1
2 BACKGROUND OF MARKET PREDICTABILITY	6
2.1 Overview	6
2.2 Market efficiency	6
2.2.1 Definition	6
2.2.2 Purely random?	8
2.3 Market inefficiency with conditional probabilities	10
2.3.1 Demonstrating market inefficiency	10
2.3.2 Existing research demonstrating predictability	12
2.3.3 Our research demonstrating predictability	16
2.3.3.1 Introduction	16
2.3.3.2 Dataset and preprocessing steps	17
2.3.3.3 Experiment 1: Test of market independence	20
2.3.3.4 Experiment 2: Escaping the bid/ask spread	24
2.3.3.5 Trends with high apparent predictability	28
2.3.4 Conclusion	30
2.4 Methods of predicting	33
2.4.1 Introduction	33
2.4.2 Fundamental and technical <i>type</i> approaches	34
2.4.2.1 Fundamental analysis	34
2.4.2.2 Technical analysis	37
2.4.2.3 Quantitative technical analysis	39
2.5 Conclusion	43
3 MACHINE LEARNING INTRODUCTION	45
3.1 Overview	45
3.2 Supervised versus unsupervised learning	45
3.3 Supervised learning algorithms	48
3.3.1 k Nearest-neighbor	48
3.3.2 Naïve Bayes	49
3.3.3 Decision table	49

3.3.4	Support Vector Machines	49
3.3.5	Artificial Neural Networks	50
3.3.6	Decision Trees	52
3.3.7	Ensembles	53
3.3.7.1	Bagging	55
3.3.7.2	Boosting	55
3.3.7.3	Combining classifiers for ensembles	56
3.4	Performance metrics	58
3.4.1	Confusion matrix and accuracy	58
3.4.2	Precision and recall	60
3.4.3	Kappa	61
3.4.4	ROC	62
3.4.5	Cost-based	64
3.4.6	Profitability of the model	65
3.5	Methods of testing	68
3.5.1	Holdout	69
3.5.2	Sliding window	71
3.5.3	Prequential	72
3.5.4	Interleaved test-then-train	72
3.5.5	k -fold cross-validation	73
3.6	Conclusion	74
4	DATA STREAM PREDICTION	76
4.1	Introduction	76
4.2	Concept drift	79
4.2.1	Definition and causes	79
4.2.2	Approaches to learning with concept drift	82
4.2.2.1	Find evidence of concept drift and then re-train	85
4.2.2.2	Assuming drift occurs	90
4.3	Adaptive models and wrapper frameworks	93
4.3.1	Adaptive Models	94
4.3.1.1	Overview	94
4.3.1.2	Existing work	95
4.3.1.2.1	Very Fast Decision Tree	95
4.3.1.2.2	Exponential fading of data	96
4.3.1.2.3	Online bagging and boosting	97
4.3.2	Wrapper Frameworks	101
4.3.2.1	Overview	101
4.3.2.2	Existing work	102
4.4	Performance and efficiency	106
4.5	Conclusion	108
5	ADDRESSING PROBLEMS SPECIFIC TO STOCK DATA	110

5.1	Imbalanced data streams	110
5.1.1	Overview	110
5.1.2	Strategies	113
5.1.2.1	Over- and under-sampling and synthetic training generation	113
5.1.2.2	Cost-based solutions	115
5.2	Preprocessing of data	117
5.2.1	Overview	117
5.2.2	<i>Bad</i> trade data and noise	118
5.2.3	Too much and too little trade data	122
5.2.4	Transformations	125
5.2.4.1	Discretization	125
5.2.4.2	Data normalization and trend correction	129
5.2.5	Attribute creation	131
5.2.5.1	Overview	131
5.2.5.2	Sentiment as indicators	132
5.2.5.3	Technical analysis indicators	134
5.2.6	Dimensionality reduction and data reduction	135
5.2.6.1	Overview	135
5.2.6.2	Filter-based feature selection	137
5.2.6.3	Wrapper feature selection	140
5.2.6.4	Embedded feature selection	144
5.2.6.5	Experiment of time complexity	144
5.3	News and its effect on price	145
5.3.1	Experiments	147
5.3.2	Discussion	150
5.4	Conclusion	150
6	OUR WRAPPER FRAMEWORK FOR THE PREDICTION OF STOCK DIRECTION	152
6.1	Overview	152
6.2	Our wrapper framework	153
6.2.1	Slow training versus fast evaluation of classifiers	156
6.2.2	Use of additional stocks	161
6.3	Benchmarks	163
6.4	Model choices	170
6.4.1	Overview	170
6.4.2	Classifiers types	170
6.4.3	Additional stocks in the classifier pool	176
6.4.3.1	Inclusion versus exclusion	176
6.4.3.2	Change in ensemble stock proportions	177
6.4.4	Feature reduction analysis	183
6.4.5	Subsets for training	190

6.4.6	Incorporating time into the predictive model	198
7	CONCLUSION AND FUTURE RESEARCH	204
7.1	Summary of prediction of stock market direction	204
7.2	Future research	208
	APPENDIX	212
A	PROBABILITY TABLES	212
B	DESCRIPTION OF DATASET	222
C	CREATION OF ATTRIBUTES WITH TECHNICAL ANALYSIS INDICATORS	231
C.1	Rate of change	232
C.2	Moving averages	233
C.2.1	Simple moving average % change	233
C.2.2	Exponential moving average % change	233
C.2.3	Exponential moving average volume weighted % change	234
C.3	Regression	234
C.4	Moving average of variance ratio	235
C.5	Relative strength index	236
C.6	Chande momentum oscillator	238
C.7	Aroon indicator	238
C.8	Bollinger Bands	240
C.9	Commodity channel index	242
C.10	Chaikin volatility	243
C.11	Chaikin money flow	244
C.12	Chaikin Accumulation/Distribution	245
C.13	Close location value	245
C.14	Moving average convergence divergence oscillator	246
C.15	Money flow index	247
C.16	Trend detection index	249
C.17	Williams %R Relative strength index	249
C.18	Stochastic Momentum Oscillator	250
C.19	Correlation analysis	251
	REFERENCES	253

LIST OF TABLES

Table

2.1	Results from t-test for the different timespans and assuming unequal variances	24
2.2	Comparing the conditional probabilities of directional movements for reversals versus continuations – brackets are the standard deviations	26
2.3	Comparing the probabilities and the level of significance for reversion-to-mean and trend continuations for a 30 second timespan	30
3.1	An example of a supervised learning dataset	46
3.2	Confusion matrix	58
3.3	Computing the Kappa statistic from the confusion matrix	62
3.4	Hypothetical cost matrix	66
3.5	Importance of using an unbiased estimate of its generalizability – trained using the dataset from Appendix B for January 3, 2012	68
3.6	Data stream with data partitioned into three subsets for cross-validation	74
4.1	Demonstrating the number of instances (minutes) until the first appearance of concept drift using Gama et al. [83] algorithm for detection of drift	90
5.1	TAQ trade data	118
5.2	Our experimental results of the Brownlees and Gallo algorithm to detect actual out-of-sequence trades	121
5.3	Time complexity (seconds) of different filter and wrapper methods . . .	145
6.1	Classifier (DT and SVM) training times (for 25 classifiers) in minutes for specific training set sizes and attribute counts	158
6.2	Benchmarks	166

6.3	Benchmarks visualized with darker shades of green representing the particular classifier is outperforming the average of all the classifiers performances on the stock and darker shades of red represents varying levels of underperformance (same as Table 6.2)	167
6.4	Average baseline classifier rank covering all 34 stocks used in the study .	169
6.5	Comparison of ensembles composed of pools comprised only of decision trees (DT), nonlinear support vector machines (SVM), artificial neural networks (ANN), equal combination of all three (DT, SVM, ANN) and a combination of the two individual best classifiers (SVM, ANN)	172
6.6	Aggregate base classifier proportion in the ensemble when base classifier was chosen by evaluating on the sliding window $t - 1$ over the length of the experiment (largest proportion in bold)	175
6.7	Including within our ensemble pool, classifiers from the stock we are predicting only (exclusion) or also adding classifiers from stocks within the same sector also (inclusion)	178
6.8	Stock n and its average proportion (over all interval) for both the pool and its selection in the ensemble	181
6.9	The number of times (out of 88 intervals) a particular stock makes up the largest proportion of the ensemble (i.e. has the most trained classifiers in the ensemble)	182
6.10	Aggregate proportion of base classifiers chosen for the ensemble trained using either the Correlation-based or Information Gain filters (base classifiers were chosen by evaluating on the sliding window $t - 1$ over the length of the experiment)	191
6.11	In minutes, average size of the training set and the average distance of the classifiers from time t for the classifiers in the pool versus the classifiers in the ensemble (larger number in bold) – test statistic at $\alpha = 0.05$	196
6.12	Comparison of classifiers from the pool versus the classifiers chosen for the ensemble	198
6.13	Including base classifiers in the pool with and without four new time attributes (best in bold) – test statistic at $\alpha = 0.05$	202
7.1	Hypothetical cost matrix	211

A.1	Conditional probabilities of market directional movements for trade-by-trade (tick) data	213
A.2	Conditional probabilities of market directional movements for 1 second timespan	214
A.3	Conditional probabilities of market directional movements for 3 second timespan	215
A.4	Conditional probabilities of market directional movements for 5 second timespan	216
A.5	Conditional probabilities of market directional movements for 10 second timespan	217
A.6	Conditional probabilities of market directional movements for 20 second timespan	218
A.7	Conditional probabilities of market directional movements for 30 second timespan	219
A.8	Conditional probabilities of market directional movements for 1 minute timespan	220
A.9	Conditional probabilities of market directional movements for 5 minute timespan	221
B.1	List of stocks used in the experiments	230

LIST OF FIGURES

Figure

2.1	Real versus random stock prices	13
2.2	Bid-ask spread schematic [192]	14
2.3	Boxplot of $\text{Pr}(+)$ for different timespans aggregated over 52 separate weeks	21
2.4	Boxplot of $\text{Pr}(+ -)$ for different timespans, along with the number of significant weeks	22
2.5	Mean conditional probabilities of depth 2 for different timespans. Until 5 to 10 seconds, predictability is higher for <i>reversals</i> of trends, after which <i>continuation</i> of trend is higher.	25
2.6	Examining monthly stability of events using 30 second interval data . . .	27
2.7	Examples of high-probability events: the trend continuation and the trend reversion-to-mean	29
2.8	Examining the daily stock price of Piedmont Natural Gas (symbol: PNY); arrows mark the dates that 10-Q (quarterly financial statements) and 10-K (annual financial statements) are released	35
2.9	The intra-day stock price of Piedmont Natural Gas (symbol: PNY) for January 23, 2012	36
2.10	The stock Exxon on January 3, 2012 with the high, low and closing prices shown on the top plot and the transaction volumes shown on the bottom plot	38
2.11	Example of a head-and-shoulders technical analysis indicator	39
2.12	Example of using Bollinger Bands as an quantitative technical analysis indicator	41
2.13	Example of using Moving Average Convergence Divergence Oscillator (MACD) as an quantitative technical analysis indicator	42

3.1	An example of an unsupervised learning technique – clustering	47
3.2	Example of a multilayer feed-forward artificial neural network	51
3.3	Artificial neural network classification error versus number of epochs . .	52
3.4	Ensemble simulation	54
3.5	ROC curve example	63
3.6	Possible directional price moves for our hypothetical example – move up, down, or no change	66
3.7	Trading algorithm process	67
3.8	Demonstrating a problem with the holdout method – averaging out pre- dictability	70
3.9	An example of the <i>sliding window</i> approach to evaluating the data stream performance	71
3.10	Data stream with data partitioned into three subsets for cross-validation	74
4.1	Naïve method of learning and testing models using stock data – using $\frac{2}{3}$ data to train and then $\frac{1}{3}$ of the data to test	78
4.2	Learning <i>instance-by-instance</i> versus by <i>chunk of instances</i>	79
4.3	Demonstrating train once and test multiple times	83
4.4	Demonstrating the loss in performance (AUC) as testing gets further away from last training data	84
4.5	Demonstrating via a stacked bar chart the change of class priors for 30 minute/instance periods for the stock Exxon for the first week of 2012 . .	86
4.6	Demonstrating the layout of our experiment using Gama et al. [83] concept Drift Detection Method (DDM)	89
4.7	Comparisons of batch and online bagging and boosting using a simulated dataset	101
4.8	Demonstrating performance (blue) and increases in training times (red) due to increases in instances	107

5.1	Demonstrating the price change of symbol XOM on January 3, 2012 . . .	112
5.2	Demonstrating IBM stock price with <i>bad</i> trade (January 03, 2005)	119
5.3	A subset of our experimental results of finding noise with the Brownlees and Gallo algorithm (noise as determined by algorithm has a green dot) .	121
5.4	Trades (transactions) often arrive at irregular time, thus causing problems when building learning algorithms	123
5.5	Demonstration of the reduction of granularity of SPY stock on January 3, 2005 from 9:30 to 10:00 a.m.	124
5.6	Open, high, low and close over a n interval period, where $n = 10$ minutes in this example	124
5.7	Symbol AXP on January 3, 2012	130
5.8	Demonstrating the “simple moving average % change” indicator	136
5.9	Three main divisions of feature selection	138
5.10	Genetic algorithm schema [244]	143
5.11	Oil services stocks reacting to the U.S. Energy Information Administration release of the petroleum status report	148
6.1	Our wrapper-based framework for the prediction of stock direction	155
6.2	Our wrapper-based framework – random starting and ending periods and random classifier types	156
6.3	Classifier training times (for 25 classifiers) – visualization of Table 6.1 . .	159
6.4	Visually demonstrating the high level of intraday correlation among 34 oil services stocks (each line represents a different normalized stock price) .	162
6.5	Visualization over 15 minutes of the changing nature of the Spearman correlation coefficient matrix over time for stocks within the same sector (oil services)	164

6.6	Comparison of ensembles composed of pools comprised only of decision trees (DT), nonlinear support vector machines (SVM), artificial neural networks (ANN), an equal combination of all three (Combined 3) and an equal combination of the two best performing base classifiers (Combined 2)	173
6.7	Process of implementing a filter-based feature subset selection procedure in our framework	185
6.8	Comparison of two different feature selection filters on the stock Exxon (symbol: XOM) using a sliding window of 1000 instances (showing only attributes 30 through 129 to save space)	187
6.9	Visualizing the different groups of attributes as a proportion of total attributes chosen by the different filter feature selection methods 46 intervals of 1000	188
6.10	Visualization of 100 base classifiers and their random start and end times over 47,000 instances	192
6.11	Each classifier is build with a training subset of size n instances and a distance (or age) of length k from the current time t	193
6.12	The moving window of size n (where n is 30,000 in our approach) limits the classifiers used in the ensemble	194
6.13	Distribution of base classifier training sets for the stock WMB (over entire experiment)	195
6.14	Demonstrating a decrease in slope before 12:30 p.m. and an increase in slope after 1:00 p.m. in the level of price moves over 0.05% (either up or down) throughout the trading day (stock: ConocoPhillips)	200
7.1	Incorrect predictions have different costs depending on the objective	209
B.1	Intraday Spearman Rank correlation over 7 months for our sector and (as a comparison) random stocks	223
B.2	January through July stock data (symbols: ANR – DVN)	224
B.3	January through July stock data (symbols: EOG – NFX)	225
B.4	January through July stock data (symbols: NOV – XOM)	226

B.5	Proportion of time defined as “move down” (red), “no change” (blue), or “move up” (green) over the course of 7 months (symbols: ANR – DVN) .	227
B.6	Proportion of time defined as “move down” (red), “no change” (blue), or “move up” (green) over the course of 7 months (symbols: EOG – NFX) .	228
B.7	Proportion of time defined as “move down” (red), “no change” (blue), or “move up” (green) over the course of 7 months (symbols: NOV – XOM) .	229
C.1	Demonstrating open, high, low, close and share volumes during an interval of 1 minute	232
C.2	Demonstrating price with SMA % change(20)	234
C.3	Price with regression(10). Red line on last price at time t represents the distance (percentage change) between the current price and the predicted.	235
C.4	Demonstrating price with MovAvgVar(5,20)	236
C.5	Demonstrating price with RSI(5)	237
C.6	Demonstrating price with CMO(5)	239
C.7	Demonstrating Aroon UpIndicator(20) and DownIndicator(20)	240
C.8	Demonstrating price with Bollinger Bands	242
C.9	Price with CCI(20)	243
C.10	Demonstrating the CLV	245
C.11	Price with DiffMACDSignal(12, 26, 9)	247
C.12	Demonstrating price with MoneyFlowIndex(15)	248
C.13	Demonstrating price with TDI(20, 20)	250
C.14	Demonstrating price with WilliamsRSI(10)	251
C.15	Demonstrating price with corrClose(3,20)	252

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.

PREVIEW