TIME SERIES PREDICTION USING RECURRENT NEURAL NETWORKS AND BOOSTING: AN EXPERIMENTAL STUDY IN PHARMACEUTICAL PRODUCT FORMULATION

by

GOH WEI YEE

Thesis submitted in fulfilment of the requirements for the degree of Master of Science

May 2002

Specially dedicated to:

my beloved parents

and all my family members

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartiest gratitude to my supervisor, Dr. Lim Chee Peng, for his guidance and motivation throughout the project. I am greatly indebted to him for his help and his precious advice towards this research. Despite his busy schedule, he is always readily reachable for the fruitful discussions. Without his invaluable collaboration, this project would not have been completed. I am greatly grateful to my co-supervisors, Dr. Peh Kok Khiang and Dr. Khairanum Subari who have given useful help and comments for my research.

I am deeply grateful to my parents, Goh Ming Tat and Wong Wong Gan, for heir love, support and caring for all these years. Their encouragements have given me trength to continue my studies to higher degrees. I am sincerely thankful to my rothers, sister and all my family members, for surrounding me with their support.

I would also like to thank my colleagues who have been generous with their ime and ideas throughout my research. I treasure their help and their support that have riven me along my research.

Last but not least, a big thank you to the Malaysian Government for providing nancial support with a special grant under the Postgraduate Scholarship Scheme.

CONTENTS

ACKNOWLEDGEMENTS ii

CONTENTS iii

LIST OF TABLES vii

LIST OF FIGURES x

ABBREVIATIONS xiii

ABSTRAK xv

ABSTRACT xvii

Chapter 1
Introduction

1.1 Preliminaries 1
1.2 Artificial Neural Networks 2
1.3 Time Series Prediction 4
1.4 Problems and Motivation 5
1.5 Research Objectives 8
1.6 Thesis Outline 10

Chapter 2
Artificial Neural Networks

2.3 Review of Artificial Neural Networks 17
2.3.1 Multi-Layer Perceptron Networks 20
2.3.2 Recurrent Neural Networks 21
2.3.2.1 Single-Layer Feedback Networks 22
2.4 The Elman Network 23
2.4.1 Dynamics of the Elman Network 25
2.4.2 The Training Algorithm 25
2.4.3 The Back-Propagation Learning 26
2.4.4 Momentum and Adaptive Learning Rate 27
2.4.5 Activation Functions 28
2.4.6 Nguyen-Widrow Initialisation 31
2.4.7 Leave-one-out Method 33
2.5 Summary 33

Chapter 3
Pharmaceutical Product Formulation: An Artificial Neural Network Approach

3.1 Introduction 35
3.2 Performance Measures 36
3.2.1 Dissolution Profiles Comparisons 36
3.2.2 Bootstrap Confidence Intervals 38
3.3 Experimental Studies 41
3.3.1 Experiment I 41
3.3.1.1 Motivation 41
3.3.1.2 Experimental Procedures 42
3.3.1.3 Results and Discussion 43
3.3.2 Experiment II 49
3.3.2.1 Motivation 49
3.3.2.2 Experimental Procedures 50
3.3.2.3 Results and Discussion 51
3.4 Summary 58

Chapter 4
An Incremental Predictor System with Boosting

4.1 Introduction 59
4.2 The Ensemble Approach 60
4.3 The AdaBoost Algorithm 64
4.4 Ensemble Size 68
4.5 Boosting and Noise 69
4.6 Modified AdaBoost 70
4.7 Simulation Studies with Benchmark Problems 73
4.7.1 The Sunspot Series 74
4.7.2 The Box-Jenkins Series 79
4.8 Summary 84

Chapter 5
The Modified AdaBoost Algorithm in Pharmaceutical Application

5.1 Introduction 87
5.2 Hypothesis Testing 87
5.3 Experimental Studies 90
5.3.1 Experiment I 90
5.3.1.1 Experimental Procedures 90
5.3.1.2 Results and Discussion 92
5.3.2 Experiment II 102
5.3.2.1 Experimental Procedures 102
5.3.2.2 Results and Discussion 103
5.4 Summary 111

Chapter 6
Conclusions and Further Work

6.1 Conclusions 113

REFERENCES 120

General References 131

APPENDICES

Appendix A

Experimental Details of Chapter 3 135

Appendix B

Benchmark Data Sets of Chapter 4 143

Appendix C

Experimental Details of Chapter 5 147

PUBLICATIONS 150

LIST OF TABLES

Table 3.1 Average difference between the test and reference dissolution profiles. 38

Table 3.2 Means of f_1 and f_2 with varying number of nodes in the hidden layer of the Elman network in Experiment I. 44

Table 3.3 Bootstrap for mean of the f_1 for 95% confidence intervals (CIs) in Experiment I with seven nodes in the hidden layer. 47

Table 3.4 Bootstrap for mean of the f_2 for 95% confidence intervals (CIs) in Experiment I with seven nodes in the hidden layer. 47

Table 3.5 Means of f_1 and f_2 with varying number of nodes in the hidden layer of the Elman network in Experiment II. 52

Table 3.6 Bootstrap for mean of the f_1 for 95% confidence intervals (CIs) in Experiment II with nine hidden nodes in the Elman network. 55

Table 3.7 Bootstrap for mean of the f_2 for 95% confidence intervals (CIs) in Experiment II with nine hidden nodes in the Elman network. 55

Table 4.1 The sum of squared errors for the single and boosted Elman networks (with 22 members) and the percentage of errors reductions using the sunspot data. 76

Table 4.2 The sum of squared errors for various models using the sunspot data. 78

Table 4.3 The mean square errors for the single and boosted Elman networks (with 18 members) and the percentage of errors reductions using the Box-Jenkins data. 81

Table 4.4 The mean square error for various models using the Box-Jenkins data. 83

Table 5.2 Comparisons of the means of f_1 and f_2 between the single and ten boosted Elman networks in Experiment I using hypotheses tests. 97

Table 5.3 Bootstrap for mean of the f_1 for 95% confidence intervals (CIs) from the boosted Elman network in Experiment I. 99

Table 5.4 Bootstrap for mean of the f_2 for 95% confidence intervals (CIs) from the boosted Elman network in Experiment I. 99

Table 5.5 Comparisons of the means of f_1 and f_2 between the single and twenty boosted Elman networks in Experiment II using hypotheses tests. 107

Table 5.6 Bootstrap for mean of the f_1 for 95% confidence intervals (CIs) from the boosted Elman network in Experiment II. 108

Table 5.7 Bootstrap for mean of the f_2 for 95% confidence intervals (CIs) from the boosted Elman network in Experiment II. 108

Appendix A

Table A1 (a)-(f) In-vitro theophylline release from pellets containing different proportions of microcrystalline cellulose (MCC) to glyceryl monostearate (GMS). 135

Table A2 The f_1 values of various proportions of MCC and GMS with different architectures of the Elman network. 137

Table A3 The f_2 values of various proportions of MCC and GMS with different architectures of the Elman network. 138

Table A4 (a)-(f) In-vitro theophylline release from pellets containing constant microcrystalline cellulose (MCC) and glyceryl monostearate (GMS) but various proportions of theophylline. 139

Table A5 The f_1 values of constant MCC and GMS but various proportions of theophylline with different architectures of the Elman network. 141

Table A6 The f_2 values of constant MCC and GMS but various proportions of theophylline with different architectures of the Elman network. 142

Appendix B

Table B1 The Sunspot data set. 143

Table B2 The Box-Jenkins data set. 145

Appendix C

Table C1 The f_1 values of various proportions of MCC and GMS using the single Elman networks and ten boosted Elman networks..... 147

Table C2 The f_2 values of various proportions of MCC and GMS using the single Elman networks and ten boosted Elman networks..... 148

Table C3 The f_1 values of various proportions of theophylline using the single Elman networks and twenty boosted Elman networks..... 148

Table C4 The f_2 values of various proportions of theophylline using the single Elman networks and twenty boosted Elman networks..... 149

LIST OF FIGURES

Figure 2.1 Processing information in an artificial neuron. 18

Figure 2.2 Representation of various artificial neural network architectures: (a) Feedforward representation; (b) Feedback representation; (c) Cellular representation. 19

Figure 2.3 The architecture of a three-layer feedforward neural network 20

Figure 2.4 The Hopfield network. 22

Figure 2.5 Architecture of the Elman network. 24

Figure 2.6 Plot of (a) the linear function and (b) its derivative. 29

Figure 2.7 Plot of (a) the sigmoid function and (b) its derivative. 29

Figure 2.8 Plot of (a) the hyperbolic tangent function and (b) its derivative. 30

Figure 3.1 Principle of the bootstrap. 40

Figure 3.2 A 3-dimensional plot of the three inputs to the Elman network in Experiment I. 46

Figure 3.3 Mean dissolution profiles of theophylline pellets with mixture of MCC and GMS in matrix ratios of (A)10:0, (B)8:2, (C)7:3, (D)6:4, (E)5:5, and (F)4:6 obtained from the Elman network (◆) and from physical experiments (■). 49

Figure 3.4 Mean of the f_1 with 95% confidence interval for various proportions of theophylline in different structures of the Elman network. 53

Figure 3.5 Mean of the f_2 with 95% confidence interval for various proportions of theophylline in different structures of the Elman network. 53

Figure 3.7 Mean dissolution profiles of formulations containing theophylline to matrix materials in ratios of (A) 6:5:5, (B) 8:5:5, (C) 10:5:5, (D) 12:5:5, (E) 14:5:5, and (F) 16:5:5 obtained from the Elman network (◆) and from physical experiments (■)..... 57

Figure 4.1 A classifier ensemble..... 62

Figure 4.2 The AdaBoost algorithm..... 66

Figure 4.3 The modified AdaBoost algorithm..... 72

Figure 4.4 Predictions of the boosted Elman network for sunspot data in the training phase and test phase, [(a), (b)] and [(c), (d)], respectively: (a) the original series (solid line) and the predicted series (dashed line); (b) the prediction errors from year 1712 to 1920; (c) the original series (solid line) and the predicted series (dashed line); (d) the prediction errors from year 1921 to 1979..... 77

Figure 4.5 Predictions of the boosted Elman network for Box-Jenkins data: (a) the original series (solid line) and the predicted series (dashed line); (b) the prediction errors of 292 samples..... 82

Figure 4.6 The mean square error curve of a single Elman network using the Box-Jenkins data..... 84

Figure 5.1 The f_1 curves of six trials of the boosted Elman networks with mixture of MCC and GMS in matrix ratios of (A)10:0, (B)8:2, (C)7:3, (D)6:4, (E)5:5, and (F)4:6..... 93

Figure 5.2 The f_2 curves of six trials of the boosted Elman networks with mixture of MCC and GMS in matrix ratios of (A)10:0, (B)8:2, (C)7:3, (D)6:4, (E)5:5, and (F)4:6..... 94

Figure 5.3 Means and 95% confidence intervals of the f_1 using 200 bootstraps from the single Elman network (dashed line) and the boosted Elman network (solid line) in Experiment I..... 100

Figure 5.4 Means and 95% confidence intervals of the f_2 using 200 bootstraps from the single Elman network (dashed line) and the boosted Elman network (solid line) in Experiment I..... 100

Figure 5.5 Mean dissolution profiles of theophylline pellets with mixture of MCC and GMS in matrix ratios of (A)10:0, (B)8:2, (C)7:3, (D)6:4, (E)5:5, and (F)4:6 obtained from the boosted Elman network (◆) and from physical experiments (■)..... 101

Figure 5.7 The f_2 curves of six trials of the boosted Elman networks with formulations containing theophylline to matrix materials in ratios of (A)6:5:5, (B)8:5:5, (C)10:5:5, (D)12:5:5, (E)14:5:5, and (F)16:5:5..... 105

Figure 5.8 Means and 95% confidence intervals of the f_1 using 200 bootstraps from the single Elman network (dashed line) and the boosted Elman network (solid line) in Experiment II. 110

Figure 5.9 Means and 95% confidence intervals of the f_2 using 200 bootstraps from the single Elman network (dashed line) and the boosted Elman network (solid line) in Experiment II. 110

Figure 5.10 Mean dissolution profiles of formulations containing theophylline to matrix materials in ratios of (A)6:5:5, (B)8:5:5, (C)10:5:5, (D)12:5:5, (E)14:5:5, and (F)16:5:5 obtained from the boosted Elman network (◆) and from physical experiments (■). 111

ABBREVIATIONS

AdaBoost	Adaptive Boosting
ANCOVA	Analysis of Covariance
ANFIS	Adaptive-Network-Based Fuzzy Inference System
ANNs	Artificial Neural Networks
ANOVA	Analysis of Variance
AR	Autoregressive
ARIMA	Autoregressive Integrated Moving Average
CIs	Confidence Intervals
FDA	United States Food and Drug Administration
FIR	Finite-Impulse-Response
FuNN	Fuzzy Neural Network
GMS	Glyceryl Monostearate
HyFIS	Hybrid Neural Fuzzy Inference System
LMS	Least Mean Square
LSP	Linear Sub-Predictor
MA	Moving Average
MCC	Microcrystalline Cellulose
MLP	Multi-Layer Perceptron
MSE	Mean Square Error

NSP	Non-Linear Sub-Predictor
OCR	Optical Character Recognition
RSM	Response Surface Methodology
SSEs	Sum of Squared Errors
UCI	University of California, Irvine

PENGGUNAAN RANGKAIAN NEURAL PERULANGAN DAN TEKNIK GALAKAN UNTUK RAMALAN SIRI MASA: SUATU KAJIAN FORMULASI PRODUK FARMASI

ABSTRAK

Tesis ini berpusat pada perkembangan teknik Rangkaian Neural Buatan (ANN) dalam menyelesaikan masalah-masalah ramalan siri masa. Penyelidikan ini tertumpu kepada penggunaan rangkaian-rangkaian neural perulangan yang menyediakan satu kerangka yang menyeluruh bagi formulasi produk farmasi melalui pendekatan ramalan siri masa. Khususnya, kerangka ini telah menjelajahi paradigma pembelajaran ANN dalam mengendalikan perancangan eksperimen dan analisis. Berdasarkan kepada kaedah-kaedah yang sedia ada, reka bentuk ANN yang baru dicadangkan untuk analisis siri masa di dalam proses formulasi produk farmasi.

Rangkaian neural perulangan jenis Elman telah digunakan untuk meramalkan profil pelarutan secara in-vitro bagi kawalan nisbah pelepasan pelet-pelet theophylline. Sambungan-sambungan perulangan di dalam rangkaian ini melibatkan penghitungan pengulangan, dan membekalkan sifat perwakilan dinamik bagi informasi dalam analisis siri masa, terutama sistem dinamik yang tidak linear. Keputusan eskperimen berjaya menunjuk potensi rangkaian-rangkaian jenis Elman di dalam formulasi produk farmasi untuk mencapai sifat kelepasan dadah yang diingini. Tambahan pula, kebolehpercayaan pencapaian rangkaian dihitungkan secara statistik dengan penggunaan kaedah ikat but.

Untuk memperbaiki ketegapan dan kebolehsesuaian rangkaian Elman, konsep daripada penggabungan keputusan alat peramal berlipat ganda untuk menghasilkan suatu kesudahan muktamad itu dicadangkan. Penubuhan teori dan strategi pengendalian bagi teknik galakan telah dihuraikan secara terperinci. Algoritma AdaBoost asal yang kerap digunakan dalam tugas klasifikasi diubahsuaikan untuk berkesesuaian di dalam rangkaian Elman berlipat ganda bagi menyelesaikan masalah ramalan siri masa. Beberapa kajian simulasi dijalankan menggunakan set-set data tanda aras, dan keputusan dibandingkan dengan pencapaian yang diperoleh dari kaedah terbitan lain. Keputusan ini telah menunjukkan bahawa rangkaian Elman berlipat ganda bersama algoritma AdaBoost ubahsuai berupaya memperbaiki anggapan umum rangkaian-rangkaian individu. Tambahan pula, masalah-masalah formulasi farmasi sebelum itu dilawat semula untuk menilai alat peramal berlipat ganda yang direka berdasarkan rangkaian Elman di dalam penggunaan praktis. Keputusan ini dinilai secara statistik dengan ujian-ujian hipotesis untuk mendapat justifikasi terhadap keberkesanan sistem yang dicadangkan. Secara keseluruhan, penyelidikan ini telah mendedahkan faedah penggunaan rangkaian Elman berlipat ganda bersama teknik galakan sebagai satu kerangka persekutuan yang mudah dan penggunaan teknik di dalam ramalan siri masa bagi formulasi produk farmasi.

ABSTRACT

This thesis is devoted to the development of Artificial Neural Network (ANN) techniques for solving time-series prediction problems. The research is focused on the use of recurrent neural networks for devising a comprehensible framework for pharmaceutical product formulation using time series prediction approach. In particular, the framework explores the learning paradigms of ANNs for conducting the experimental design and analysis. Based upon existing methodologies, novel ANN architectures are proposed for time series analyses in the process of pharmaceutical product formulation.

The Elman recurrent neural network is employed for the prediction of in-vitro dissolution profiles of matrix-controlled-release theophylline pellets preparations. Feedback links in this network perform recursive computation, and provide the ability of dynamical representation of information in time series analyses, especially for non-linear dynamical systems. The experimental results have successfully demonstrated the potentials of the Elman-based networks for formulating pharmaceutical products to meet the desired drug release characteristics. Furthermore, reliability of the network performance is statistically assessed using the bootstrap method.

The theoretical foundations and operational strategies of *boosting* have been elaborated in details. The standard AdaBoost algorithm that is often employed in classification tasks is modified to accommodate multiple Elman networks for time-series prediction problems. Several simulation studies are conducted using benchmark data sets, and the results are compared with those obtained from other published methods. The results indicate that multiple Elman networks coupled with the modified AdaBoost algorithm are capable of improving generalisation of individual networks. In addition, the pharmaceutical formulation problems are re-visited to assess the practical applicability of multiple predictors devised based on the Elman networks. The results are evaluated statistically using hypotheses tests to justify the effectiveness of the proposed system. In summary, this research work has revealed the benefits of using multiple Elman-based networks with boosting as a unified and convenient framework for utilizing techniques in time series prediction for pharmaceutical product formulation.

Chapter 1

Introduction

1.1 Preliminaries

Time series prediction is a common problem in many fields. Economists want to forecast economic and financial information based on current and historical measurements. Demographers want to predict changes in population in the future. In the prediction of physical time series, meteorologists want to predict weather and temperature. A consumer products company likes to analyse marketing data, *viz.* the export totals in successive months, profits in successive years, and growth in sales for a new product. It is indeed important to forecast future sales so as to plan production. In process control, performance of a manufacturing process needs to be monitored for statistical quality control. Therefore, time series prediction is an area that drives researchers to seek underlying principles through prediction to explain behaviour of the observed systems. Since accurate forecasts are required in so many management and engineering decision-making problems, it is hence worthwhile to conduct research into the aspects of time series prediction.

In recent years, work on Artificial Neural Networks (ANNs) has become increasingly popular. An ANN is a massively parallel, adaptive dynamical system with self-learning capability that is capable of performing useful information-processing

capabilities of generalisation, non-linearity, adaptivity, and input-output mapping have been the important properties for the analysis of time series prediction. Therefore, this thesis is concerned with the investigation of ANNs in real world applications, particularly in pharmaceutical product formulation, using time series prediction approach.

Based upon the existing boosting theory (Freund & Schapire, 1997), a novel ANN system with ensemble has been proposed to improve the predictive accuracy and generalisation capability of individual ANNs. Simulation of the resulting system has been performed both on benchmark data sets as well as pharmaceutical databases comprising drug dissolution profiles of theophylline formulations.

In the following sections, an introduction to ANNs and time series prediction are described. Difficulties of classical statistical approaches to time series prediction, and motivation of the ANN approach are discussed in this chapter. Then, the research objectives are defined, and an overview of organisation of this thesis is presented.

1.2 Artificial Neural Networks

The more recent history of ANN systems began with the work by McCulloch and Pitts (1943) who studied mathematical models of the brain. However, Hebb (1949) was one of the first to suggest the idea of learning in ANNs through adapting the connections between nodes or processing elements. The connectionist architecture was used to represent "*knowledge*". Even today, still, many of the learning models are some

ANN system was discussed. Amongst the earlier models, Rosenblatt (1958) had developed several variations of networks called *perceptrons* and had studied different forms of learning. The basic perceptron network was a threshold logic unit made up of three layers: an input sensory layer that was randomly connected to an association layer that was, in turn, connected to an output response or classification layer. If the cumulative inputs from the sensory layer to the association layer exceeded some threshold, that unit fired an impulse to activate the response layer and produced an output of +1, and produced an output of 0, if not. During the early 1970s, researchers such as Kohonen (1972) and Anderson (1972) had conducted investigations on *associative memories* in order to explore their computational power and limitations. According to Anderson (1972), a highly simplified associative ANN model was one group of neurons projected to another group of neurons. Anderson (1972) assumed that the activity level of a neuron in the output layer was simply a weighted sum of the activity levels of the neurons in the input layer, with the synaptic weights according to the Hebbian learning rule. It followed by Hopfield (1982) that introduced an associative memory network called Hopfield network. One of the most important developments of recent ANN research is the discovery of a *supervised learning algorithm* to adjust weights in multi-layer feedforward networks. The algorithm is known as *back-propagation* since the weights are adjusted from the output layer backwards layer-by-layer to reduce the output errors (Rumelhart *et al.*, 1986a).

Although researchers attempt to emulate the structure and function of biological neurons, artificial neuron models are, still, not exactly constrained by real neurons and are based loosely on biology. People hardly understand the behaviour of real nervous

designed in order to realise the specific computational problems and their architectures are based upon the problem to be solved. As a problem-solving tool, ANNs have found promising results in various disciplines of science and engineering.

1.3 Time Series Prediction

A time series is a chronological sequence of observations on a particular variable (Bowerman & O'Connell, 1979). Broadly speaking, a time series is a collection of observations made sequentially in time (Chatfield, 1996). It is said to be *continuous* when observations are made continuously in time. In contrast, it is said to be *discrete* when observations are taken only at specific time intervals, usually equally spaced (Chatfield, 1996). The term "discrete" is compactable even when the measured variable in certain series is a continuous variable. Discrete time series, called the *sampled* series, can arise in several ways: digitise a continuous time series at some intervals of time, or aggregate (accumulate) the values over equal intervals of time. The special feature of time series analysis is the fact that successive observations are usually *not* independent and that the analysis must take into account the time *order* of the observations. Indeed, future values may be predicted from past observations. If a time series can be predicted exactly by some mathematical function, it is said to be *deterministic*. Unfortunately, exact predictions are not always possible because of some unknown factors. Thus, future values are normally predicted by having a probability distribution that is conditioned by the knowledge of past values (Chatfield, 1996). Such model is called the *stochastic* model. In a stochastic process, the underlying probability mechanism will vary with time. A simplification, called *stationary* model, is introduced

constant mean level (Yaffee & McGee, 2000). However, many time series in real applications are often better represented as *non-stationary* that have no natural mean.

1.4 *Problems and Motivation*

Time series consist of several components, *i.e.* trend, cycle, seasonal variations and irregular fluctuations (Bowerman & O'Connell, 1979). Normally, linear and periodic components (*i.e.* trends, cyclical and seasonal variations) are easy to model and to remove from the time series using traditional methods, *e.g.* regression analysis, exponential smoothing, decomposition, and Box-Jenkins. One of the famous analyses, the regression analysis, is normally used in time series forecasting (Chatfield, 1996). The regression analysis is a study of relationships among variables. It uses observations of the studied variables to calculate a curve of best fit so that the behaviour of the variables can be estimated and predicted. The regression analysis can be seen as a case of function approximation. In Caswell (1982), several methods of regression analysis, *i.e.* the three-point method, least squares method, and method of moving averages, have been described for detecting the trend of a time series. However, several disadvantages occur. For the three-point and least squares methods, a lot of information is lost in the averaging process. Although the three-point method is easy to execute, it always results in a linear trend that may not be appropriate. The method of moving averages suffers from limitation that trend values are not available at the beginning and the end of the series. Even though the regression analysis associated with mathematical models has been applied to predict non-linear time series models, this method fails to predict models that are highly non-linear. Thus, this method hardly fits the considering model