

**RARE EVENT FORECASTING BY USING HIGHLY
NON-LINEAR NEURAL NETWORKS AND DATA MINING
KERNELS: CASE STUDY OF EARTHQUAKE PREDICTION**

by

Zhou Nannan



Master of Science in E-Commerce Technology

2013



**Faculty of Science and Technology
University of Macau**

Rare Event Forecasting by Using Highly Non-linear Neural
Networks and Data Mining Kernels: Case Study of Earthquake
Prediction

by
Zhou Nannan

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science in E-Commerce Technology

Faculty of Science and Technology
University of Macau

2013

Approved by _____
Supervisor

Date _____

In presenting this thesis in partial fulfillment of the requirements for a Master's degree at the University of Macau, I agree that the Library and the Faculty of Science and Technology shall make its copies freely available for inspection. However, reproduction of this thesis for any purposes or by any means shall not be allowed without my written permission. Authorization is sought by contacting the author at

Address: University of Macau

Telephone:

E-mail: nannan-z@hotmail.com.cn



Signature_____

Date_____

University of Macau

Abstract

RARE EVENT FORECASTING BY USING HIGHLY NON-LINEAR NEURAL
NETWORKS AND DATA MINING KERNELS: CASE STUDY OF
EARTHQUAKE PREDICTION

By Zhou Nannan

Thesis Supervisor: Dr. Simon Fong
Master of Science in E-Commerce Technology

Earthquake forecasting (prediction) is known to be a challenging research program (problem) for which no single prediction method can claim to be the best and is hard to obtain the high and consistent prediction accuracy in worldwide. At large, earthquake data when viewed as a time series over a long time, exhibits a complex pattern that is composed of a mix of statistical features. This is due to the fact that the magnitude variation is high; hence the white noise component is dominating. A single prediction algorithm often does not yield an optimal forecast by analyzing over a long series that is composed of a large variety of features. In the past many attempts by using traditional time series forecasting algorithms have been made, and the prediction models are usually based on the input of univariate time series. In this thesis, a new analytic framework is proposed that allows these mixed features from the time series to be automatically extracted by a computer program, and fed into a decision tree classifier for choosing an appropriate method for the current forecasting task. The motivation behind this concept of a new analytic framework is to let the data decide which prediction algorithm should be adopted, adaptively across different periods of the time series. With a case study of global maximal earthquake prediction by different time series, the performance results obtained by traditional univariate time series forecasting methods, GMDH¹ alone, and extended GMDH with residual-feedback are generated and compared. The experiment results show that GMDH with residual-feedback can yield the lowest error of all relatively. It is

¹ Group Method of Data Handling was applied in a great variety of areas for data mining and knowledge discovery, forecasting and systems modeling, optimization and pattern recognition.

forecasted that in merely one year from now, an earthquake of magnitude 10.2136 in Richter scales will strike the world, at the MAPE (mean absolute percentage error) of 2.211%. This case study is a novel approach that converts univariate time series to multivariate data is proposed, namely residual-feedback, for providing relevant multivariate inputs to GMDH or polynomial neural networks. This is important because the strength of GMDH just like any neural network is on predicting outcomes from multivariate data. GMDH is a well-known ensemble type of prediction method that is capable of modelling highly non-linear relations, and achieving an optimal accuracy by inducing through all possible structures of polynomial forecasting (prediction) models. The contributions of this thesis are: (1) a framework of automatic forecasting which is very suitable for real-time earthquake monitor is proposed, (2) an investigation on how different features of the data series are coupled with different prediction algorithms for the best possible accuracy, and (3) the case study of global maximal earthquake event prediction by different time series period by GMDH methods is significant and will potentially make a high impact to the development of forecasting techniques. In lieu of earthquake prediction, the same residual-feedback GMDH model could be used to forecast (predict) any kind of time series speculation with very low error measures.

Keywords: Earthquake prediction, Time series forecasting, Forecasts and predictions, Automatic and adaptive forecasting, ARIMA, Holt-Winter's, GMDH, Neural network, Data pre-processing.

TABLE OF CONTENTS

List of Figures	iii
List of Tables	vi
List of Abbreviations	vii
APPENDIX.....	viii
Chapter 1: Introuduction.....	1
1.1 Background.....	1
1.2 Motivations	3
1.3 Summary of Data Mining	4
1.3.1 Definitions of the Data Mining.....	5
1.3.2 Process of Data Mining.....	6
1.3.3 Task and Function of Data Mining.....	8
1.4 Data Mining in Earthquake Prediction	13
1.4.1 Major Methods of the Earthquake Prediction.....	13
1.4.2 Earthquake Data and Characteristic.....	14
1.4.3 Introducing Data Mining Methods.....	16
1.5 Orgnization of the Dissertation.....	17
Chapter 2: Related Earthquake Works.....	18
2.1 Main Earthquake Precursors.....	19
2.1.1 Strange Animal Behavior.....	21
2.1.2 Frequency/Density Patterns in Seismic Data.....	23
2.1.3 Geophysical and Geochemical Anomalies	24
2.1.4 EM and Ionospheric Anomalies.....	27
2.1.5 Thermal Anomalies.....	28
2.2 Temporal and Spatial Data Mining.....	29
2.3 Model Selection	31
2.4 Parameter Selection	32
2.5 Data Selection	33

Chapter 3: Framework for Forecasting	34
3.1 Methodology	34
3.2 Streaming Operation	35
3.3 Decomposing the Time Series into Salient Features	43
3.4 Decision Tree As Forecaster Selector	44
Chapter 4: Experiments	46
4.1 A Case Study of Time Series Forecasting by Residual-Feedback	48
4.2 Theoretical Framework	51
4.2.1 COMBI Algorithm	53
4.2.2 MIA Algorithm	55
4.3 Extended GMDH with Additional Inputs	56
4.3.1 Residual Feedback GMDH	60
4.3.2 Our Proposed Methodology	62
4.4. Our Experiment	64
4.5 Experiment Results and Discussion	69
4.5.1 Experiment Results	69
4.5.2 Discussion	73
Chapter 5: Conclusion	88
Bibliography	91
APPENDIX (A-F)	104

LIST OF FIGURES

Figure 1: Data mining as the process or steps of knowledge discovery by Usama Fayyad.....	6
Figure 2: Data mining as the process or steps of knowledge discovery by Jiawei Hart	7
Figure 3: The pictures of legends about earthquake on different places	20
Figure 4: Adaptive forecasting methodology	34
Figure 5: Operational view of adaptive forecasting process.....	35
Figure 6: Examples of other analysis by using the statistical characteristics of Earthquake time series	36
Figure 7: The contrastive plot of M-t ($M \geq 8.0$) plot and sunspot number line plot	39
Figure 8: The contrastive plot of F-t ($M \geq 8.0$) plot and sunspot number line plot.....	39
Figure 9: The plot of arisen earthquake ($M \geq 8.0$) year	40
Figure 10: Decision tree for deciding which forecasting method to use, based on the salient characteristics of the time series that are currently under observation.....	45
Figure 11: (a) ACF Plot of the earthquake frequency time series;(b) PACF Plot of the earthquake frequency time series.....	47
Figure 12: Examples of forecasting models by different techniques that illustrate the concept of “better fit does not always mean better prediction!”	50
Figure 13: Data structure of GMDH process.....	53
Figure 14: Two calculation methods of Hurst exponent series.....	59
Figure 15: A fragment of global great earthquake of RF-type of neural network.....	62
Figure 16: Contour visual diagram of three variables, Year, Magnitude and Residual.....	63
Figure 17: The Operational Flow of the Residual-Feedback GMDH Method	64
Figure 18: Performance comparison of different GMDH methods in MAPE.....	70
Figure 19: Performance segment comparison of different GMDH methods in MAPE	71

Figure 20: Comparison of magnitude time series, R1 and R2	71
Figure 21: Comparison of magnitude time series, R1 and R2 in normalize.....	72
Figure 22: Performance comparison of different time series prediction methods in MAPE in weka.....	74
Figure 23: The algorithm of Classifiers.functions.MultilayerPerceptron	75
Figure 24: The algorithm of Classifiers.meta.AdditiveRegression	75
Figure 25: The algorithm of Classifiers.meta.RegressionByDiscretization	75
Figure 26: The algorithm of Classifiers.rules.DecisionTable	76
Figure 27: The example graphics of RF-GMDH results	78
Figure 28: GMDH COMBI on the univariate time series with no feedback.	84
Figure 29: GMDH COMBI on the multivariate time series with R1 feedback.	84
Figure 30: GMDH COMBI on the multivariate time series with R2 feedback.	84
Figure 31: GMDH COMBI on the multivariate time series with H feedback.....	84
Figure 32: GMDH MIA on the univariate time series with no feedback.	84
Figure 33: GMDH MIA on the multivariate time series with R1 feedback.	84
Figure 34: GMDH MIA on the multivariate time series with R2 feedback.	84
Figure 35: GMDH MIA on the multivariate time series with H feedback.	84
Figure 36: GMDH COMBI on the multivariate time series with R1 and R2 feedback.....	85
Figure 37: GMDH COMBI on the multivariate time series with R1 and H feedback.....	85
Figure 38: GMDH COMBI on the multivariate time series with R2 and H feedback.....	85
Figure 39: GMDH COMBI on the multivariate time series with R1, R2 and H feedback.....	85
Figure 40: GMDH MIA on the multivariate time series with R1 and R2 feedback.....	85
Figure 41: GMDH MIA on the multivariate time series with R1 and H feedback.....	85
Figure 42: GMDH MIA on the multivariate time series with R2 and H feedback.....	85
Figure 43: GMDH MIA on the multivariate time series with R1, R2 and H feedback.....	85

Figure 44: The best result of GMDH MIA on the multivariate time series with R1,
R2 and H feedback in Ya'an.....87



LIST OF TABLES

Table 1: A part of examples of global great earthquake data	16
Table 2: Example of global Great earthquake data from 1700-2010.....	38
Table 3: The vocabulary of sunspot number.....	41
Table 4: The phase of global great earthquake	41
Table 5: The statistics of earthquake number of eight phases	42
Table 6: Some useful salient features for our model	43
Table 7: Comparison of adaptive model and others alone.....	48
Table 8: Scale of Hurst (H) Exponent Classifications and Continuity Intensity	60
Table 9: Acronyms (abbreviation) of various combinations of datasets	68
Table 10: Performance comparison of different GMDH methods in error measures and model quality: Earthquake Data Yearly	69
Table 11: Output comparison of different GMDH methods in residual errors and prediction results for earthquake data	70
Table 12: Performance comparison of different time series prediction methods in weka in error measures and model quality: Earthquake Data Yearly.....	74
Table 13: Output comparison of different time series prediction methods in weka in residual errors and prediction results for earthquake data	77
Table 14: Illustrate tp, fp, fn and tn	80
Table 15: Comparison of magnitude time series, R1, R2 and H in statistics	81
Table 16: Output comparison of different time series prediction methods in residual errors for earthquake data in different Cycle of Time.....	82
Table 17: Output comparison of different time series prediction results for earthquake data in different Cycle of Time	83
Table 18: Output comparison of different time series prediction methods in Precision, Recall and F1 score for earthquake data in different Cycle of Time	83
Table 19: Output comparison of different best methods prediction results for Sichuan Ya'an.....	86

LIST OF ABBREVIATIONS

KDD	Knowledge Discovery in Database
ACM	Association for Computing Machinery
KD	Knowledge Discovery
TFT	Traditional Forecasting Techniques
USGS	U.S. Geological Survey
CENC	China Earthquake Networks Center
AEIC	Alaska Earthquake Information Center
NGDC	National Geophysical Data Center
NOAA	National Oceanic and Atmospheric Administration
AR	Autoregressive
MA	Moving Average
ARIMA	Autoregressive Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
ARIMAX	AutoRegressive Integrated Moving Average with eXtra
GMDH	Group Method Data Handling
COMBI-GMDH	Combinatorial Group Method Data Handling
MIA-GMDH	Multilayer Iterative Group Method Data Handling
RF	Residual Feedback
LAG-1	One-step Lag
EEG	Electroencephalography
U-TS	Univariate Time Series
M-TS	Multivariate Time Series
MAPE	Mean Absolute Percentage Error

MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
NMAE	Normalized Mean Absolute Error
NRMSE	Normalized Root Mean Squared Error
RMSPE	Root Mean Square Prediction Error
CI	Criterion Index
DES	Double Exponential Smoothing
EM	Electromagnetic
GPS	Global Position System
SES	Seismic Electric Signals
ULF	Ultra Low Frequency
GIS	Geographical Information Systems
TEC	Total Electron Content
UIIF	Ultra High Frequency
GRL	Geophysical Research Letters

APPENDIX

APPENDIX A: Autoregressive Model Selection Criteria. (Source: [119])

APPENDIX B: The output diagrams of model fitting and forecast results by half year.

APPENDIX C: The output diagrams of model fitting and forecast results by quarter.

APPENDIX D: The output diagrams of model fitting and forecast results by month.

APPENDIX E: The current model panel of our COMBI model on components, coefficients and formula by year.

APPENDIX F: The current model panel of our MIA model on components, coefficients and formula by year.



ACKNOWLEDGMENTS

I am very grateful to my thesis supervisors, Dr. Simon Fong, who gave me a lot of encouragement and support. He always gave me valuable guidance and suggestions all the times. Especially, something happened in my home about my mother accident. During that time, the teacher Dr. Simon Fong can help and understand me. His support has meant a lot to me during this difficult time.

Finally, I would like to acknowledge my family and friends' support during my thesis research. Hope my mother recovers soon. Everything is okay in the end.

