การพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ โดยใช้วิธีการเติมเต็มแบบจัดกลุ่มข้อมูลให้สมบูรณ์และวิธีประสานผลของตัวแบบโครงข่ายประสาท

นางสิรภัทร เชี่ยวชาญวัฒนา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-17-6750-1

INCOMPLETE TIME-SERIES DATA FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE AND ENSEMBLING NEURAL NETWORK MODEL

Mrs. Sirapat Chiewchanwattana

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Science
Department of Mathematics
Faculty of Science
Chulalongkorn University
Academic year 2005

ISBN 974-17-6750-1

Thesis Title	INCOMPLETE TIME-SERIES DATA FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE AND ENSEMBLING NEURAL NETWORK MODEL
By	Mrs. Sirapat Chiewchanwattana
Field of Study	Computer Science
Thesis Advisor	Professor Chidchanok Lursinsap, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctor's Degree

.... Dean of the Faculty of Science

(Professor Piamsak Menasveta, Ph.D.)

THESIS COMMITTEE

.... Chairman

(Associate Professor Jack Asavanant, Ph.D.)

.... Thesis Advisor

(Professor Chidchanok Lursinsap, Ph.D.)

.... Member

(Associate Professor Boonserm Kijsirikul, Ph.D.)

.... Member

(Assistant Professor Rajalida Lipikorn, Ph.D.)

.... Member

(Chularat Tanprasert, Ph.D.)

สิรภัทร เชี่ยวชาญวัฒนา : การพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ โดยใช้วิธีการเติมเต็มแบบจัดกลุ่มข้อมูลให้สมบูรณ์ และวิธีประสานผลของตัวแบบโครงข่ายประสาท : (INCOMPLETE TIME-SERIES DATA FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE AND ENSEMBLING NEURAL NETWORK MODEL). อ. ที่ปรึกษา : ศาสตราจารย์ ดร. ชิดชนก เหลือสินทรัพย์, 133 หน้า. ISBN 974-17-6750-1.

วิทยานิพนธ์นี้นำเสนอ การพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ โดยอาศัยการจำลองรูปแบบของโครงข่ายประสาทเทียม ซึ่งการจำลองนั้นสามารถแบ่งได้เป็นสองขั้นตอนดังนี้ ขั้นตอนที่หนึ่ง ทำการเติมเต็มข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์นั้นให้สมบูรณ์ ในขั้นตอนที่สองทำการพยากรณ์ข้อมูลอนุกรมเวลาที่ได้จากขั้นตอนที่หนึ่ง การแก้ปัญหาในงานนี้คือพัฒนาแบบจำลองโครงข่ายประสาทเทียมใหม่ สำหรับการพยากรณ์ข้อมูลอนุกรมเวลาที่ไม่สมบูรณ์ และยังต้องสามารถให้ความถูกต้องในการพยากรณ์เพิ่มขึ้นด้วย โดยได้นำเสนอแบบจำลองโครงข่ายประสาทเทียม สองแบบ แบบแรก ใช้วิธีการเติมเต็มข้อมูลแบบ EM หลายลักษณะ และวิธีการเติมเต็มข้อมูลแบบ Spline ซึ่งข้อมูลหลายๆ ชุดที่ถูกเติมเต็มจากหลายๆ วิธีนั้น จะถูกนำมาสอนโดยใช้โครงข่ายประสาทเทียม MLP โดยใช้แบบขยาย Kalman Filtering จากนั้นทำการประสานผลลัพธ์ของโครงข่ายประสาทเทียมทุกโครงข่ายเข้าด้วยกัน แบบจำลองโครงข่ายนี้ให้ชื่อว่า โครงข่าย FI-GEM แบบที่สองปรับเปลี่ยนมาใช้โครงข่ายประสาทเทียม FIR เพื่อทำการพยากรณ์ จากนั้นผลลัพธ์ของโครงข่ายประสาทเทียมทุกโครงข่ายจะถูกประสานเข้าด้วยกันโดยใช้วิธีการเลือกโครงข่ายแบบ genetic algorithm ให้ชื่อแบบจำลองโครงข่ายนี้ว่า โครงข่าย RMD-FSE นอกจากนั้นยังได้นำเสนอวิธีการเติมเต็มข้อมูลแบบใหม่ เพื่อปรับปรุงการประมาณค่าข้อมูลที่หายไปนั้นให้ได้ค่าที่ถูกต้องมากยิ่งขึ้น โดยได้ใช้เทคนิคการจัดกลุ่ม โดยอาศัยคุณลักษณะของรูปแบบข้อมูลที่มีอยู่จริง แนวคิดหลักคือ ทำการตัดแบ่งข้อมูลอนุกรมเวลาออกเป็นหลายๆ ชิ้นที่มีขนาดต่างๆ กัน วิธีการคำนวณหาค่าข้อมูลที่หายไป จะคำนวณหาจากชิ้นข้อมูลที่มีความคล้ายกับชิ้นที่มีข้อมูลที่หายไปมากที่สุด แล้วทำการคำนวณหาค่าข้อมูลที่หายไปนั้น ให้ชื่อว่าขั้นตอนวิธี WDC ซึ่งสามารถให้ผลที่เทียบเท่าหรือดีกว่าวิธีอื่น เช่น EM, MI, OCSFCM และ Spline ในกรณีของข้อมูลอนุกรมเวลาที่ไม่คงที่

ภาควิชา	**คณิตศาสตร์**	ลายมือชื่อนิสิต	สิ.ภัพ
สาขาวิชา	**วิทยาการคอมพิวเตอร์**	ลายมือชื่ออาจารย์ที่ปรึกษา	C.L.
ปีการศึกษา	2548		

4373850523 : MAJOR COMPUTER SCIENCE

KEY WORD: INCOMPLETE TIME-SERIES PREDICTION /MISSING DATA /CLUSTERING FILL-IN TECHNIQUE / ENSEMBLING NEURAL NETWORK.

SIRAPAT CHIEWCHANWATTANA: INCOMPLETE TIME-SERIES DATA FORECASTING BASED ON CLUSTERING FILL-IN TECHNIQUE AND ENSEMBLING NEURAL NETWORK MODEL. THESIS ADVISOR: PROF. CHIDCHANOK LURSINSAP, Ph.D., 133 pp. ISBN 974-17-6750-1.

This dissertation demonstrates the problem of incomplete time-series prediction by modelling the forecasting of several natural and social phenomena. The modeling consists of two main steps. The first step is to estimate the collected incomplete data, which are considered as missing data or missing values. The second step is to predict new data based on the nature of the data obtained from the first step. Our solution is to develop a new neural network model for forecasting incomplete time-series data and improving the accuracy of prediction. Two neural network models are proposed. First, various versions of EM-based algorithm and smoothing spline interpolation are used to preprocess the incomplete data sets. The individual networks are trained by supervised multilayer perceptron(MLP) with extended Kalman filtering. The ensemble construction is used for the combination of the individual networks. We name this type of network Fill-In - Generalized Ensemble Method (FI-GEM) networks. Second, each individual network uses a Finite Impulse Response model to perform the prediction. The outputs of all individual neural networks are combined by the genetic algorithm-based selective neural network ensemble method (GASEN). We denote this network as a reconstructed missing data-finite impulse response selective ensemble (RMD-FSE) network. Moreover, we proposed a new fill-in technique that is improved for estimating missing values based on clustering technique for characterizing the pattern of incomplete time-series data. The main idea is the time-series data are divided into separate subsequences of different sizes and, therefore, each subsequence can be viewed as a window. The imputation of missing samples is achieved by finding a complete subsequence similar to the missing sample subsequence and imputing the missing samples from this complete subsequence.

The imputation accuracy of the proposed algorithm, namely varied window clustering (WDC) algorithm is comparable or better than the others traditional methods such as: the spline interpolation, the multiple imputation (MI), and the optimal completion strategy fuzzy c-means algorithm (OCSFCM) in case of the non-stationary time-series data.

Department	**Mathematics**	Student's signature..........................
Field of study	**Computer Science**	Advisor's signature..........................
Academic year	2005	

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Chidchanok Lursinsap, for his contributions to all aspects of my work on this thesis and encouragement and for giving lots of his precious time to supervise this research work. There is one more unofficial supervisor for this thesis: Professor Chee Hung Henry Chu of the CACS, ULL. I thank him both for the close collaboration on this thesis. His sincere help and guidance have been extreamly helpful to me.

The financial support of the ministry of university affairs of Thailand and in part by the National Electronic and Computer Technology Center (NECTEC), the National Science and Technology Development Agency (NSTDA) are also gratefully acknowledged. I would like to thank the thesis committee for their helpful comments and suggestions.

Special thanks are also expressed to Dr. Khamron Sunat, Professor Suchada Siripant and Sub.Lt. Supannika Lursinsap for giving me the encouragement and support. I am extreamly grateful to my family, especially my parents, for their love, patience, and countless encouraging words.

Tables of Contents

Chapter **page**

Chapter **page**

List of Tables

List of Figures

Figure **page**

Figure **page**

Figure **page**

Figure **page**

Figure **page**

Figure **page**

List of Abbreviations

FI-GEM	Fill-In - Generalized Ensemble Method
GEM	Generalized Ensemble Method
RMD-FSE	Reconstructed Missing Data-Finite Impulse Response Selective Ensemble
WDC	Varied Window Clustering
MLP	Multi-Layer Perceptron
ML	Maximum Likelihood
EM	Expectation and Maximization
MI	Multiple Imputation
REG-EM	Regularized Expectation and Maximization
MAR	Missing at Random
OCSFCM	Optimal Completion Strategy Fuzzy C-Mean
MSE	Mean Square Error
CORR	Pearson's Correlation
GASEN	Genetic Algorithm-based Selective Neural Network Ensemble

List of Symbols

x_t	Value of a System at Time t
P_d	Performance Index
P'_d	Average Performance Index
E_i^T	Mean Square Error of the Tested Network of Run i
E_j^{RN}	Mean Square Error of the Reference Network of Run j
E_j^{RF}	Mean Square Error of Proposed Network of Run j
R	Number of Runs per Fill-In Method
L	Number of Different Percentage of Missing Data
M	Number of Missing Value
MSE	Mean Squared Error
x	Actual Values
$\widehat{x}$	Prediction Values
CORR	Pearson's correlation
N_{ij}^{Actual}	the Normalized Zero Mean of the Actual Values
$N_{ij}^{Predict}$	the Normalized Zero Mean of the Estimated Values
P_{Imp}	the Average Imputation Performance Index
f_{GEM}	Outputs of All Individual Neural Networks
$f_i(\mathbf{x})$	Output Value of Network i
α_i	Weighting Parameter for Network i
C_{ij}	Elements of the Covariance Matrix of the Errors from the Function Estimators f_i and f_j

k	Window Size of k
$o_j^{(q)}$	Output Signal of j^{th} Neuron in the q^{th} Layer
$w_{i,j}^{(q)}$	Connection Weight coming from the i^{th} Neuron in the $(q-1)$ Layer to the j^{th} Neuron in the q^{th} layer
δ	Similarity Correlation in terms of Cosine
K	Length of the Subsequences
β	Distance between two Subsequences
$\mathbf{v}_\tau$	Target Subsequence
$\mathbf{v}_q$	Reference Subsequence
$x_m^{(\mathbf{v}_\tau)}$	the Considered Missing Value at Time m of the Target Subsequence $\mathbf{v}_\tau$
$x_m^{(\mathbf{v}_q)}$	Value at Time m of the Reference Subsequence $\mathbf{v}_q$
θ_l	Left Difference between $x_{m-1}^{(\mathbf{v}_q)}$ and $x_{m-1}^{(\mathbf{v}_\tau)}$
θ_r	Right Difference between $x_{m+1}^{(\mathbf{v}_q)}$ and $x_{m+1}^{(\mathbf{v}_\tau)}$
$\hat{x}_m^{(\mathbf{v}_\tau)}$	Estimated Value of $x_m^{(\mathbf{v}_\tau)}$
$G_{(d\%)}$	the Goodness of Partitioning Window Size at $d\%$ of Missing
κ_j	Window Size at Order j
q_{κ_j}	the number of points in window of size κ_j