

การเรียนรู้แบบโยนแล้วทิ้งสำหรับการจำแนกประเภทกลุ่มข้อมูลที่เข้าที่ละชุดโดยการประยุกต์  
เวอร์เซทส์อัลลิบติกเบซิสฟังก์ชันในการคำนวณแบบทีละกลุ่มข้อมูล



นายเปรม จันทรสว่าง



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการ

คอมพิวเตอร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2556

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

DATA-THROWAWAY LEARNING FOR STREAMING CHUNK DATA CLASSIFICATION BY  
APPLYING A VERSATILE ELLIPTIC BASIS FUNCTION (VEBF) TO SINGLE-CLASS-WISE  
COMPUTATION

Mr. Prem Junsawang

A Dissertation Submitted in Partial Fulfillment of the Requirements  
for the Degree of Doctor of Philosophy Program in Computer Science and  
Information Technology

Department of Mathematics and Computer Science

Faculty of Science

Chulalongkorn University

Academic Year 2013

Copyright of Chulalongkorn University



5 1 7 3 9 1 0 6 2 3

Thesis Title DATA-THROWAWAY LEARNING FOR STREAMING  
CHUNK DATA CLASSIFICATION BY APPLYING A  
VERSATILE ELLIPTIC BASIS FUNCTION (VEBF) TO  
SINGLE-CLASS-WISE COMPUTATION

By Mr. Prem Junsawang

Field of Study Computer Science and Information Technology

Thesis Advisor Assistant Professor Suphakant Phimoltares, Ph.D.

Thesis Co-Advisor Professor Chidchanok Lursinsap, Ph.D.

---

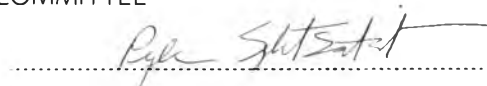
Accepted by the Faculty of Science, Chulalongkorn University in Partial  
Fulfillment of the Requirements for the Doctoral Degree



..... Dean of the Faculty of Science

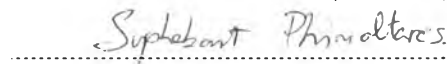
(Professor Supot Hannongbua, Dr.rer.nat.)

#### THESIS COMMITTEE



..... Chairman

(Associate Professor Peraphon Sophatsathit, Ph.D.)



..... Thesis Advisor

(Assistant Professor Suphakant Phimoltares, Ph.D.)



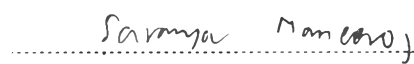
..... Thesis Co-Advisor

(Professor Chidchanok Lursinsap, Ph.D.)




..... Examiner

(Professor Boonserm Kijirikul, Ph.D.)



..... Examiner

(Assistant Professor Saranya Maneeroj, Ph.D.)



..... External Examiner

(Chularat Tanprasert, Ph.D.)

เปรม จันทรสว่าง : การเรียนรู้แบบโยนแล้วทิ้งสำหรับการจำแนกประเภทกลุ่มข้อมูลที่เข้าที่ละชุดโดยการประยุกต์เวอร์เซโทลล์อิลลิปติกเบซิสฟังก์ชันในการคำนวณแบบทีละกลุ่มข้อมูล. (DATA-THROWAWAY LEARNING FOR STREAMING CHUNK DATA CLASSIFICATION BY APPLYING A VERSATILE ELLIPTIC BASIS FUNCTION (VEBF) TO SINGLE-CLASS-WISE COMPUTATION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. ศุภกานต์ พิมลธเรศ, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ศ. ดร. ชิตชนก เหลือสินทรัพย์, 67 หน้า.

ปัญหาการจำแนกประเภทกลุ่มข้อมูลขนาดใหญ่และข้อมูลที่เข้าที่ละชุดถือว่าเป็นปัญหาที่น่าสนใจและท้าทายซึ่งจะพบได้ในการประยุกต์ใช้งานจริง เช่น ข้อมูลด้านการเงิน ข้อมูลการวินิจฉัยทางการแพทย์ งานทางด้านความรู้จำรูปแบบและเหมืองข้อมูล โดยในหลายกรณีการที่จะได้มาซึ่งฐานข้อมูลที่สมบูรณ์สำหรับนำมาใช้เป็นข้อมูลสำหรับการสร้างตัวแบบในการจำแนกนั้นเป็นไปได้ยาก ดังนั้นในงานนี้ ผู้วิจัยนำเสนอวิธีการเรียนรู้แบบโยนแล้วทิ้งสำหรับการจำแนกประเภทกลุ่มข้อมูลที่เข้าที่ละชุด (Data-throwaway Learning for Streaming Chunk, DLSC) โดยการประยุกต์เวอร์เซโทลล์อิลลิปติกเบซิสฟังก์ชัน (Versatile Elliptic Basis Function, VEBF) ในการคำนวณแบบทีละกลุ่มข้อมูล ซึ่งแนวคิดของวิธีการเรียนรู้ที่นำเสนอนี้อาศัยหลักการการเรียนรู้แบบเพิ่มขึ้นและการเรียนรู้แบบที่ข้อมูลที่ใช้ในการเรียนรู้เพียงครั้งเดียว ในงานนี้ ผู้วิจัยดำเนินการทดลองในลักษณะตามรูปแบบของข้อมูลชุดสอนที่ได้มา ประกอบด้วยกรณีข้อมูลชุดสอนที่สมบูรณ์และข้อมูลชุดสอนที่เข้ามาทีละชุด โดยทำการเปรียบเทียบผลของวิธีการเรียนรู้ที่นำเสนอกับกลุ่มวิธีการเรียนรู้ทั้งแบบการเรียนรู้แบบแบทช์และการเรียนรู้แบบเพิ่มขึ้น และชุดข้อมูลที่นำมาใช้ทดสอบประสิทธิภาพนั้นมีความหลายหลายทั้งในแง่ของจำนวนข้อมูลซึ่งมีค่าตั้งแต่ 150 ถึง 581,012 ข้อมูลและจำนวนคุณลักษณะประจำตั้งแต่ 4 ถึง 1,558 คุณลักษณะ จากผลการทดลองสรุปได้ว่าขั้นตอนวิธีการเรียนรู้ที่นำเสนอให้ค่าความถูกต้องสูงสุดในหลายๆกรณี นอกจากนี้ยังใช้เวลาในการเรียนรู้ จำนวนโหนดในชั้นซ่อน และความยืดหยุ่นของโครงสร้างของตัวแบบที่ดีกว่าวิธีการที่นำมาเปรียบเทียบ วิธีการที่นำเสนอสามารถจัดการกับปัญหาการจำแนกข้อมูลที่มีขนาดใหญ่และข้อมูลที่เข้ามาทีละชุดได้

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์

สาขาวิชา วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ

ปีการศึกษา 2556

ลายมือชื่อนิสิต Prem Junsawang

ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก Suphant Pimlathors

ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม C. Ly



# # 5173910623 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY  
 KEYWORDS: CLASSIFICATION / INCREMENTAL LEARNING ALGORITHM / VERSATILE  
 ELLIPTIC BASIS FUNCTION / STREAMING CHUNK DATA

PREM JUNSAWANG: DATA-THROWAWAY LEARNING FOR STREAMING  
 CHUNK DATA CLASSIFICATION BY APPLYING A VERSATILE ELLIPTIC BASIS  
 FUNCTION (VEBF) TO SINGLE-CLASS-WISE COMPUTATION. ADVISOR: ASST.  
 PROF. SUPHAKANT PHIMOLTARES, Ph.D., CO-ADVISOR: PROF. CHIDCHANOK  
 LURSINSAP, Ph.D., 67 pp.

Recently, the large data and streaming chunk data classification problems are the interesting and challenging problems in many real world applications such as finance, medical diagnosis, pattern recognition, and data mining. In most cases, a complete set of database for building a classifier is not provided in advance. In this work, the Data-throwaway Learning for Streaming Chunk data classification (DLSC) by applying a Versatile Elliptic Basis Function (VEBF) to single-class-wise computation is proposed. The proposed learning method is based on incremental learning and one-pass-throwaway learning concepts. In this work, the experiment is conducted in two scenarios based on the pattern of given training data including complete training data and streaming training data. The experimental results of the proposed method are compared with both of batch learning and incremental learning algorithms on various data sets with different sizes from 150 to 581,012 samples and attributes from 4 to 1,558. The experimental results show that the DLSC yields the highest classification accuracies in most cases with faster incremental learning, fewer number of used hidden neurons and more flexible structure than the compared methods. The proposed method is suitable for coping with big data classification problem and handling streaming data as well.

Department: Mathematics and  
 Computer Science  
 Field of Study: Computer Science and  
 Information Technology

Student's Signature *Prem Junsawang*  
 Advisor's Signature *Suphakant Phimoltares*  
 Co-Advisor's Signature *Chidchanok Lursinsap*

Academic Year: 2013



## ACKNOWLEDGEMENTS

This dissertation would not have been possible without the guidance and the helps of several individuals with their valuable assistance in the preparation and completion of this study.

Firstly, I would like to express my deepest appreciation to Assistant Professor Dr. Suphakant Phimoltares and Professor Dr. Chidchanok Lursinsap, who are my advisor and co-advisor respectively, for his valuable advice, guidance in this dissertation. It is honor to work with them.

I also would like to thank the chairman and member of the committee of this research work namely Associate Professor Dr. Peraphon Sophatsathit, Professor Dr. Boonserm Kitsirikul, Assistant Professor Dr. Saranya Maneeroj and Dr. Chularat Tanprasert, for their valuable comments and personal efforts in reviewing this dissertation.

I am grateful to Advance Virtual and Intelligence Computing (AVIC) Research Center for facilitating me to achieve this work. I am also grateful to the Development and Promotion of Science and Technology Talents Project (DPST) for financial support.

I would like to thank Dr. Peerapol Khunarsal, Dr. Sirilak Areerachakul, friends and everyone at AVIC Research Center for a very nice cheer up and helpful suggestion.

Finally, I would not be able to accomplish this dissertation without the loves and supports from my beloved parents with the invaluable source of strengths.



## CONTENTS

|  | Page |
|--|------|
| THAI ABSTRACT .....  | iv   |
| ENGLISH ABSTRACT .....   | v    |
| ACKNOWLEDGEMENTS .....   | vi   |
| CONTENTS .....   | vii  |
| LIST OF TABLES .....   | ix   |
| LIST OF FIGURES .....  | x    |
| CHAPTER I INTRODUCTION.....  | 1    |
| 1.1. Statement of the Problem .....  | 1    |
| 1.2. Related Works.....  | 2    |
| 1.3. Objectives.....   | 5    |
| 1.4. Scope of Work.....  | 5    |
| CHAPTER II VERSATILE ELLIPTIC BASIS FUNCTION NEURAL NETWORK (VEBFNN) .....       | 6    |
| 2.1 Versatile Elliptic Basis Function (VEBF) .....                               | 6    |
| 2.2 Structure of Versatile Elliptic Basis Function Neural Network (VEBFNN) ..... | 7    |
| 2.2 Orthonormal Basis Computation .....  | 9    |
| 2.2.1 Principal Component Analysis (PCA).....                                    | 10   |
| 2.2.2 Orthonormal basis vectors by PCA .....                                     | 12   |
| 2.3 VEBF Learning Algorithm for one incoming datum.....                          | 13   |
| CHAPTER III PROPOSED METHODOLOGY .....   | 16   |
| 3.1 Parameters Update.....   | 18   |
| 3.1.1 Center vector and covariance matrix update .....                           | 18   |
| 3.1.2 Width vector update .....  | 19   |
| 3.1.3 Merge Criterion .....  | 22   |
| 3.2 The Proposed Learning Algorithm.....   | 23   |
| 3.2.1 Data-throwsaway Learning Streaming Chunk (DLSC) algorithm .....            | 25   |
| CHAPTER IV EXPERIMENTS AND RESULTS.....  | 32   |
| 4.1 Experiments for Complete Training Data Scenario .....                        | 32   |



|  |    |
|--|----|
| 4.1.1 Experimental setting for complete training data scenario .....   | 34 |
| 4.1.2 Experimental results for complete training data scenario .....   | 36 |
| 4.1.3 Comparison Results on Effect of Order of Presented Classes .....   | 41 |
| 4.1.4 Influence of center selection and initial width parameter on classification accuracy and the number of hidden neurons of DLSC..... | 46 |
| 4.2 Experiments for Streaming Training Data Scenario .....   | 52 |
| 4.2.1 Experimental setting for streaming training data scenario .....  | 53 |
| 4.2.2 Experimental results for streaming training data scenario .....  | 55 |
| CHAPTER V CONCLUSION AND DISCUSSION.....   | 60 |
| REFERENCES .....   | 62 |
| APPENDIX.....  | 65 |
| VITA.....  | 68 |





## LIST OF TABLES

|   |    |
|---|----|
| Table 1: Description of each data set for complete training data scenario.....  | 33 |
| Table 2: Parameter setting in each data set for complete training data scenario .....   | 35 |
| Table 3: Comparison results of average accuracy with standard deviation ( $\bar{x} \pm sd$ ) for complete training data scenario.....   | 37 |
| Table 4: Comparison results of the average number of hidden neurons or prototypes with standard deviation ( $\bar{x} \pm sd$ ) for complete training data scenario.....                   | 40 |
| Table 5: The ratio of hidden neurons or prototypes of each method with respect to that of DLSC.....   | 41 |
| Table 6: Comparison results of the average learning time (s) with standard deviation ( $\bar{x} \pm sd$ ) for complete training data scenario.....  | 42 |
| Table 7: Ratio of learning time of each method with respect to that of DLSC.....  | 43 |
| Table 8: Comparison results of average accuracy (%) and standard deviation ( $\bar{x} \pm sd$ ) on holdout validation of ten distinctive orders presented patterns.....                   | 44 |
| Table 9: Comparison results of number of hidden neurons or prototypes and standard deviation ( $\bar{x} \pm sd$ ) on holdout validation of ten distinctive orders presented patterns..... | 45 |
| Table 10: Description of each data set for streaming training data scenario.....  | 52 |
| Table 11: Parameter setting in each data set for streaming training data scenario .....   | 54 |
| Table 12: Comparison results of average with standard deviation ( $\bar{x} \pm sd$ ) of classification accuracy on eleven data sets.....  | 57 |
| Table 13: Comparison results of average with standard deviation ( $\bar{x} \pm sd$ ) of number of hidden neurons on eleven data sets.....   | 58 |
| Table 14: Comparison results of average with standard deviation ( $\bar{x} \pm sd$ ) of learning time with standard deviation on eleven data sets.....                                    | 59 |

## LIST OF FIGURES

|   |    |
|---|----|
| Figure 1: The structure of versatile elliptic basis function neural network.....  | 8  |
| Figure 2: Example of streaming chunk data classification in two-dimensional space. ....   | 16 |
| Figure 3: Example of class-wise streaming chunk in two-dimensional space.....   | 18 |
| Figure 4: An example of how to cover new incoming data by updating VEBF<br>parameters.....  | 21 |
| Figure 5: Two cases of merge process. ....  | 22 |
| Figure 6: Example of the proposed method in two-dimensional space. ....   | 24 |
| Figure 7: The average of standard deviation values of classification accuracy (Acc.)<br>and the number of hidden neurons (Neu.) for ten distinctive presented<br>orders ..... | 46 |
| Figure 8: Influence of different initial center vectors on classification accuracy (%) ...  | 48 |
| Figure 9: Influence of ten different initial center vectors of DLSC method .....  | 49 |
| Figure 10: Influence of the predefined $\delta$ setting versus accuracy (%) .....   | 50 |
| Figure 11: Influence of the predefined $\delta$ setting versus the number of hidden<br>neurons.....   | 51 |