# เมตริกเทียมสำหรับการจำแนกประเภทข้อมูลอนุกรมเวลาโดยใช้อัลกอริทึมเนียเรสท์เนเบอร์

นายธีศิษฏ์ กอศรีลบุตร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

# PSEUDOMETRICS FOR TIME SERIES DATA CLASSIFICATION USING THE NEAREST NEIGHBOR ALGORITHM

Mr. Teesid Korsrilabutr

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2007

| | |
|---|---|
| Thesis Title | PSEUDOMETRICS FOR TIME SERIES DATA CLASSIFICATION USING THE NEAREST NEIGHBOR ALGORITHM |
| By | Mr. Teesid Korsrilabutr |
| Field of Study | Computer Engineering |
| Thesis Advisor | Professor Boonserm Kijsirikul, Ph.D. |

---

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

..... Dean of the Faculty of Engineering
(Associate Professor Boonsom Lerdhirunwong, Dr.Ing.)

THESIS COMMITTEE

..... Chairman
(Assistant Professor Athasit Surarerks, Ph.D.)

..... Thesis Advisor
(Professor Boonserm Kijsirikul, Ph.D.)

..... Member
(Chotirat Ratanamahatana, Ph.D.)

..... Member
(Associate Professor Thanaruk Theeramunkong, Ph.D.)

ธีศิษฏ์ กอศรีลบุตร : เมตริกเทียมสำหรับการจำแนกประเภทข้อมูลอนุกรมเวลาโดยใช้อัลกอริทึมเนียเรสท์เนเบอร์ (PSEUDOMETRICS FOR TIME SERIES DATA CLASSIFICATION USING THE NEAREST NEIGHBOR ALGORITHM) อ.ที่ปรึกษา : ศ.ดร. บุญเสริม กิจศิริกุล, 54 หน้า.

เรานำเสนอว่าเมตริกเทียม ซึ่งเป็นมาตรวัดระยะห่างที่ซับแอดดิทีฟ มีคุณสมบัติเพียงพอที่จะเป็นโครงสร้างที่ดีสำหรับกระทำการจำแนกแบบอย่างด้วยเนียเรสท์เนเบอร์ มีผลทางทฤษฎีจำนวนหนึ่งที่รับประกันความถูกต้องของเคเนียเรสท์เนเบอร์แบบเชิงเส้นกำกับ เมื่อจำนวนตัวอย่างสอนโตขึ้น ผลเหล่านี้เป็นจริงภายใต้เงื่อนไขว่า มาตรวัดระยะห่างเป็นเมตริก ผลเหล่านี้ยังเป็นจริงสำหรับเมตริกเทียมหลังจากการปรับแต่งทางเทคนิค ณ ขณะนี้ยังไม่มีคำตอบว่าผลเหล่านี้สามารถเป็นจริงสำหรับมาตรวัดระยะห่างที่ไม่ซับแอดดิทีฟหรือไม่ เมตริกเทียมยังมีประโยชน์ในทางปฏิบัติอีกด้วย หากเรามีมาตรวัดระยะห่างที่ซับแอดดิทีฟแล้ว มาตรวัดนั้นจะมีความได้เปรียบมาตรวัดที่ไม่ซับแอดดิทีฟอย่างน้อยหนึ่งประการ กล่าวคือ เราสามารถนำมาตรวัดดังกล่าวไปใช้ได้โดยตรงในระบบที่ใช้ประโยชน์จากคุณสมบัติซับแอดดิทีฟเพื่อทำให้การค้นหาเพื่อนบ้านใกล้ที่สุดเร็วขึ้น

งานนี้มุ่งเน้นที่เมตริกเทียมสำหรับอนุกรมเวลา เราเสนอสองกรอบงานเพื่อศึกษาและออกแบบมาตรวัดระยะห่างซับแอดดิทีฟและตัวอย่างจำนวนหนึ่งของมาตรวัดระยะห่างที่ออกแบบได้จากกรอบงานทั้งคู่ กรอบงานหนึ่งมีความทั่วไปสูงกว่าอีกหนึ่งกรอบงานและสามารถใช้เพื่อปรับแต่งมาตรวัดจากอีกกรอบงานหนึ่งเพื่อให้ประสิทธิภาพในการจำแนกประเภทสูงขึ้นได้ ผลการทดลองจำแนกประเภทโดยใช้เนียเรสท์เนเบอร์ร่วมกับฟังก์ชันที่ออกแบบได้ เปรียบเทียบกับมาตรวัดระยะห่างที่เป็นที่รู้จักดี รวมถึง Dynamic Time Warping แสดงให้เห็นว่ามาตรวัดระยะห่างที่ออกแบบขึ้นใช้งานได้จริงสำหรับการจำแนกประเภทข้อมูลอนุกรมเวลา

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา .....

ปีการศึกษา 2550

##4870673721 : MAJOR COMPUTER ENGINEERING
KEYWORD : SEQUENCE / TIME SERIES / METRIC

TEESID KORSRILABUTR : PSEUDOMETRICS FOR TIME SERIES DATA CLASSIFICATION USING THE NEAREST NEIGHBOR ALGORITHM. THESIS ADVISOR : Prof. Boonserm Kijsirikul, Ph.D., 54 pp.

We propose that pseudometric, a subadditive distance measure, has sufficient properties to be a good structure to perform nearest neighbor pattern classification. There exist some theoretical results that asymptotically guarantee the classification accuracy of $k$-nearest neighbor when the sample size grows larger. These results hold true under the assumption that the distance measure is a metric. The results still hold for pseudometrics up to some technicality. Whether the results are valid for the non-subadditive distance measures is still left unanswered. Pseudometric is also practically appealing. Once we have a subadditive distance measure, the measure will have at least one significant advantage over the non-subadditive; one can directly plug such distance measure into systems which exploit the subadditivity to perform faster nearest neighbor search techniques.

This work focuses on pseudometrics for time series. We propose two frameworks for studying and designing subadditive distance measures and a few examples of distance measures resulting from the frameworks. One framework is more general than the other and can be used to tailor distances from the other framework to gain better classification performance. Experimental results of nearest neighbor classification of the designed pseudometrics in comparison with well-known existing distance measures including Dynamic Time Warping showed that the designed distance measures are practical for time series classification.

| | | |
|---|---|---|
| Department | Computer Engineering | Student's signature Teesid Korsrilabutr |
| Field of study | Computer Engineering | Advisor's signature ..... |
| Academic year | 2007 | |

# Acknowledgments

I love my parents. Thank heaven there is the word "parents" in this language so that I need to put neither of them first in the sentence, which might imply that I love either one more than the other. Fortunately I have only one sibling, so I have no problem mentioning my sister that I have the same kind of feeling for her. My grandma and aunties are also unforgettable as well.

My gratitude to:
my advisor, for everything that merits acknowledgments, including his quick and useful advice, he knew when was the right time my work was matured enough for publications and when it was not. Vit Niennattrakul, for the talk about his work, that must be the first time I knew there are people that do $k$-NN with non $\ell^p$ metrics; Prasertsak Pungprasertying, for encouraging me by saying [tɕěŋ] (in IPA notation, literally "cool") after being the first to read the first draft of the early version of my work, and for his pointer I always used during presentations; Ratthachat Chatpatanasiri, for suggesting me to try taking courses at the Department of Mathematics; Pasakorn Tangchanachaianan, for his insight that I had not utilized the full potential of dynamic programming; Tanasanee Phienthrakul, for making me realized that my presentation could have been well extended pass an hour long if I were not careful; MIND lab members, for being my audiences; Acting Second Lieutenant Tanawut Wuttanaprechakit, for lending me his laptop during my presentation; the committee, for building the cozy atmosphere before my oral defense, it is still unknown to me how they managed to do so despite having me felt obliged to to end my presentation in 20 minutes, their questions, discussions, and suggestions has led to some improvements in the thesis.

My effort to be frugal was assisted by the `pdftoppm` and `compare` programs from Xpdf and ImageMagick software suites. I thank their developers for helping me identify pages that need reprints after revisions, and for half the lab's ink, papers, and time saved.

This is the only page I have to use special vertical spacing to fit more text. Nonetheless, should I be grateful for any other entities or events, they can put what they like below.

⋮

# Contents

Page

# List of Tables

# List of Figures