

เทคนิคการแตกครึ่งตามสารสนเทศสำหรับซีพอร์ทเวกเตอร์แมชชีนแบบหลายประเภท



นางสาวปทุมศิริ สงศิริ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2548

ISBN 974-53-2565-1

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

AN INFORMATION-BASED DICHOTOMIZATION TECHNIQUE
FOR MULTICLASS SUPPORT VECTOR MACHINES

Miss Patoomsiri Songsiri

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

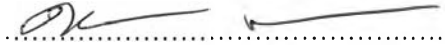
Chulalongkorn University

Academic Year 2005

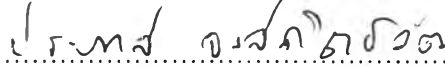
ISBN 974-53-2565-1

หัวข้อวิทยานิพนธ์	เทคนิคการแตกครึ่งตามสารสนเทศ สำหรับซอฟต์แวร์เมทริกซ์แบบหลายประเภท
โดย	นางสาวปทุมศิริ สงศิริ
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล
อาจารย์ที่ปรึกษาร่วม	ดร.ฐิมาพร เพชรแก้ว

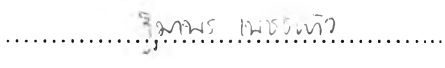
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์ฉบับนี้
เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต



..... คณะบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์


..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.ประภาส จงสิตต์ยวัฒน์)


..... อาจารย์ที่ปรึกษา
(รองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล)


..... อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร.ฐิมาพร เพชรแก้ว)


..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ้มปิยะกรณ์)

ปทุมศิริ สงศิริ : เทคนิคการแตกครึ่งตามสารสนเทศสำหรับซัพพอร์ตเวกเตอร์แมชชีนแบบหลายประเภท. (AN INFORMATION-BASED DICHOTOMIZATION TECHNIQUE FOR MULTICLASS SUPPORT VECTOR MACHINES) อ.ที่ปรึกษา: รศ.ดร.บุญเสริม กิจศิริกุล, อ. ที่ปรึกษาร่วม: อ.ดร.ฐิมาพร เพชรแก้ว, 74 หน้า. ISBN 974-53-2565-1.

การแก้ปัญหาการจำแนกแบบหลายประเภทด้วยซัพพอร์ตเวกเตอร์แมชชีนโดยส่วนใหญ่จะพิจารณาเป็นปัญหาของการนำตัวจำแนกแบบสองประเภทหลายตัวมาใช้ร่วมกัน วิธีการดังกล่าวยังมีข้อจำกัดในเรื่องของความถูกต้องและจำนวนครั้งในการจำแนก งานวิจัยนี้นำเสนอวิธีการใหม่ในการจำแนกข้อมูลด้วยเทคนิคการแตกครึ่งตามสารสนเทศโดยสร้างต้นไม้สำหรับการจำแนกแบบไบนารีจากตัวจำแนกแบบสองประเภท ซึ่งแต่ละโนดของต้นไม้จะเป็นตัวจำแนกแบบสองประเภทที่มีค่าเอนโทรปีต่ำที่สุด วิธีนี้สามารถลดจำนวนตัวจำแนกแบบสองประเภทที่ใช้ในการจำแนกได้เป็นลอการิทึมของจำนวนประเภทซึ่งต่ำกว่าวิธีอื่น จากผลการทดลองแสดงให้เห็นว่าวิธีนี้สามารถลดจำนวนครั้งของการจำแนกได้ และยังคงให้ค่าความถูกต้องใกล้เคียงกับวิธีอื่น

ภาควิชา..... วิศวกรรมคอมพิวเตอร์.....ลายมือชื่อนิสิต..... *Janis* *ณัฐ*
 สาขาวิชา.....วิทยาศาสตร์คอมพิวเตอร์...ลายมือชื่ออาจารย์ที่ปรึกษา..... *ดร. ก.*
 ปีการศึกษา ...2548.....ลายมือชื่ออาจารย์ที่ปรึกษาร่วม..... *ฐิมาพร (เพชรแก้ว)*

46703630 : MAJOR COMPUTER SCIENCE

KEY WORD: INFORMATION-BASED DICHOTOMIZATION / MULTICLASS SUPPORT VECTOR MACHINES

PATOOMSIRI SONGSIRI: AN INFORMATION-BASED DICHOTOMIZATION TECHNIQUE FOR MULTICLASS SUPPORT VECTOR MACHINES. THESIS ADVISOR: ASSOC. PROF. BOONSERM KIJSIRIKUL, Ph.D., THESIS COADVISOR : THIMAPORN PHETKAEW,Ph.D., 74 pp. ISBN 974-53-2565-1.

Approaches for solving a multiclass classification problem by Support Vector Machines (SVMs) are typically to consider the problem as combination of two-class classification problems. Previous approaches have some limitations in classification accuracy and the number of evaluations. This research proposes a novel method that employs information-based dichotomization for constructing a binary classification tree. Each node of the tree is a binary SVM with the minimum entropy. Our method can reduce the number of binary SVMs used in the classification to the logarithm of the number of classes which is lower than previous methods. The experimental results show that the proposed method takes lower the number of evaluations while it maintains accuracy compared to other methods.

Department..... Computer Engineering...Student's..... Patomsiri Songsiri
Field of study.....Computer Science.....Advisor's..... Boonserm Kijirikul
Academic year ...2005.....Coadvisor's..... Thimaporn Phetkaew

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์อย่างยิ่งของอาจารย์ที่ปรึกษารองศาสตราจารย์ ดร.บุญเสริม กิจศิริกุล และอาจารย์ที่ปรึกษาร่วม อาจารย์ ดร.ฐิมาพร เพชรแก้ว ซึ่งท่านได้ให้ความรู้ คำแนะนำ และข้อคิดเห็นอันเป็นประโยชน์ต่องานวิจัย จนทำให้การวิจัยในครั้งนี้สำเร็จออกมาด้วยดี

ขอขอบคุณรองศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา และผู้ช่วยศาสตราจารย์ ดร.ญาใจ ลิ้มปิยะกรณ์ กรรมการสอบวิทยานิพนธ์ที่กรุณาเสียสละเวลา ให้คำแนะนำ ตรวจสอบ และแก้ไขวิทยานิพนธ์ฉบับนี้

ขอบคุณสมาชิกห้องปฏิบัติการ Machine Intelligence & Knowledge Discovery (MIND-LAB) และห้องปฏิบัติการอื่นๆ ที่คอยช่วยเหลือให้คำแนะนำและให้กำลังใจด้วยดีเสมอมา

ท้ายที่สุด ผู้วิจัยใคร่ขอกราบขอบพระคุณบิดา มารดา รวมถึงทุกคนในครอบครัวที่ให้กำลังใจและให้การสนับสนุนเรื่อยมาจนสำเร็จการศึกษา

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช
สารบัญภาพ	ญ
สารบัญตาราง.....	ฎ
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของการวิจัย	2
1.3 ขอบเขตของการวิจัย.....	3
1.4 ขั้นตอนและวิธีดำเนินการวิจัย.....	3
1.5 ประโยชน์ที่คาดว่าจะได้รับ	3
1.6 ลำดับการจัดเรียงเนื้อหาในวิทยานิพนธ์.....	4
1.7 ผลงานที่ตีพิมพ์จากวิทยานิพนธ์.....	4
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง.....	5
2.1 ทฤษฎีที่เกี่ยวข้อง.....	5
2.1.1 ซัพพอร์ตเวกเตอร์แมชชีน	5
2.1.1.1 ซัพพอร์ตเวกเตอร์แมชชีนแบบเชิงเส้น	5
2.1.1.2 ซัพพอร์ตเวกเตอร์แมชชีนแบบไม่เชิงเส้น	8
2.1.2 ทฤษฎีสารสนเทศ	9
2.2 งานวิจัยที่เกี่ยวข้อง	10
2.2.1 วิธีจำแนกแบบหนึ่งต่อที่เหลือ	10
2.2.2 วิธีแมชชีน	12
2.2.3 วิธีดีดีเอจี	13
2.2.4 วิธีเอดีเอจี	14
2.2.5 วิธีอาร์เอดีเอจี.....	15
2.2.6 วิธีจำแนกข้อมูลแบบแตกครึ่งแบบสมดุล	16

บทที่ 3 เทคนิคการจำแนกข้อมูลแบบแตกครึ่งตามสารสนเทศ.....	20
3.1 แนวคิดเบื้องต้นในการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภท	20
3.2 ขั้นตอนในการสร้างต้นไม้สำหรับการจำแนกแบบหลายประเภทด้วยการแตกครึ่งตามสารสนเทศ	22
3.3 การตัดเล็มและการกำหนดขอบเขตความผิดพลาด	27
บทที่ 4 การทดลองและผลการทดลอง.....	31
4.1 ชุดข้อมูลที่ใช้ในการทดลอง	31
4.2 วิธีการทดลอง.....	33
4.3 ผลการทดลอง	34
4.3.1 ชุดข้อมูล Glass.....	35
4.3.2 ชุดข้อมูล Satimage	37
4.3.3 ชุดข้อมูล Segment	39
4.3.4 ชุดข้อมูล Shuttle.....	41
4.3.5 ชุดข้อมูล Vowel	43
4.3.6 ชุดข้อมูล Soybean.....	45
4.3.7 ชุดข้อมูล Letter.....	47
4.3.8 ชุดข้อมูล Isolet	49
4.3.9 ชุดข้อมูล Thaiprinted Character 1	51
4.3.10 ชุดข้อมูล Thaiprinted Character 2.....	53
4.4 เวลาที่ใช้ในการสอน.....	54
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ	58
5.1 สรุปผลการวิจัย	58
5.2 ข้อเสนอแนะ.....	59
รายการอ้างอิง	61
ภาคผนวก	62
ภาคผนวก ก	63
ภาคผนวก ข	72
ประวัติผู้เขียนวิทยานิพนธ์	74

สารบัญภาพ

	หน้า
รูปที่ 1 ตัวอย่างของการจำแนกข้อมูลสองประเภทด้วยซัพพอร์ทเวกเตอร์แมชชีน.....	6
รูปที่ 2 แนวคิดการแมปข้อมูลแบบเป็นไม่เชิงเส้นไปสู่ปริภูมิอันดับสูง	8
รูปที่ 3 ความสัมพันธ์ระหว่างความน่าจะเป็นและค่าสารสนเทศ.....	9
รูปที่ 4 ตัวอย่างการสร้างระนาบของการจำแนกแบบหนึ่งต่อที่เหลือ สำหรับปัญหา 4 ประเภท	11
รูปที่ 5 วิธีแมกซิมสำหรับปัญหา 4 ประเภท	12
รูปที่ 6 โครงสร้างของดีดีเอจี้สำหรับปัญหา 4 ประเภท	13
รูปที่ 7 โครงสร้างของเอดีเอจี้สำหรับปัญหา 8 ประเภท	14
รูปที่ 8 โครงสร้างการจำแนกข้อมูลของอาร์เอดีเอจี้	15
รูปที่ 9 ตัวอย่างตำแหน่งของข้อมูลกรณีปัญหา 4 ประเภท	16
รูปที่ 10 การแบ่งข้อมูลด้วยตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้ กรณีปัญหา 4 ประเภท.....	17
รูปที่ 11 โครงสร้างการจำแนกข้อมูลของการแตกครึ่งแบบสมดุลงกรณีปัญหา 4 ประเภท	18
รูปที่ 12 (ก) การแทนอักขระด้วยต้นไม้แบบไบนารี กรณีที่ความน่าจะเป็นในการเกิดของแต่ละ อักขระเท่ากัน.....	20
รูปที่ 12 (ข) การแทนอักขระด้วยต้นไม้แบบไบนารี กรณีที่ความน่าจะเป็นในการเกิดของแต่ละ อักขระไม่เท่ากัน.....	20
รูปที่ 13 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภท ของปัญหา 4 ประเภท.....	22
รูปที่ 14 การแบ่งข้อมูลด้วยตัวจำแนกแบบสองประเภททุกแบบที่เป็นไปได้ กรณีปัญหา 4 ประเภท.....	23
รูปที่ 15 ต้นไม้สำหรับการจำแนกแบบหลายประเภทที่สร้างจากค่าเอนโทรปี กรณีดีที่สุด สำหรับปัญหา 4 ประเภท.....	25
รูปที่ 16 ตัวอย่างระนาบแบ่งประเภทข้อมูลที่ทำให้ตำแหน่งของข้อมูลประเภทเดียวกันตกอยู่ ทั้งสองด้านของระนาบ.....	26
รูปที่ 17 (ก) การตัดเล็มกรณีค่าร้อยละของการตัดเล็มเป็น 10	27
รูปที่ 17 (ข) การตัดเล็มกรณีค่าร้อยละของการตัดเล็มเป็น 40	27
รูปที่ 18 ตัวอย่างช่วงค่าประสิทธิภาพโดยนัยทั่วไปของตัวจำแนกข้อมูล 325 ตัว	28
รูปที่ 19 การตรวจสอบไขว้ 5 พับ	31

รูปที่ 20 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Glass	34
รูปที่ 21 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Satimage	36
รูปที่ 22 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Segment	38
รูปที่ 23 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Shuttle.....	40
รูปที่ 24 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Vowel	42
รูปที่ 25 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Soybean	44
รูปที่ 26 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Letter	46
รูปที่ 27 ความน่าจะเป็นในการเกิดข้อมูลแต่ละประเภทของชุดข้อมูล Isolet	48
รูปที่ 28 การแบ่งข้อมูลด้วยตัวจำแนกแบบสองประเภทของตัวจำแนก 1-3 และ 1-4.....	58

สารบัญตาราง

หน้า

ตารางที่ 1 : ข้อมูลที่ใช้ในการทดลองจาก UCI Machine Learning Repository และ Thai printed character recognition dataset	30
ตารางที่ 2(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Glass	34
ตารางที่ 2(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Glass.....	35
ตารางที่ 2(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Glass	35
ตารางที่ 3(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Satimage	36
ตารางที่ 3(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Satimage... 37	
ตารางที่ 3(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Satimage	37
ตารางที่ 4(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Segment	38
ตารางที่ 4(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Segment .39	
ตารางที่ 4(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Segment.....	39
ตารางที่ 5(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Shuttle.....	40
ตารางที่ 5(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Shuttle	41
ตารางที่ 5(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Shuttle	41
ตารางที่ 6(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Vowel.....	42
ตารางที่ 6(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Vowel	43
ตารางที่ 6(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Vowel	43
ตารางที่ 7(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Soybean	44
ตารางที่ 7(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Soybean .. 45	
ตารางที่ 7(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Soybean	45
ตารางที่ 8(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Letter	46
ตารางที่ 8(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Letter.....	47
ตารางที่ 8(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Letter	47
ตารางที่ 9(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Isolet	48
ตารางที่ 9(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Isolet	49
ตารางที่ 9(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Isolet	49
ตารางที่ 10(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Thaiprinted Character 1 และ 2	50

ตารางที่ 10(ข) : การเปรียบเทียบจำนวนครั้งในการจำแนกที่คาดหวังของชุดข้อมูล Thaiprinted Character 1	51
ตารางที่ 10(ค) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Thaiprinted Character 1	51
ตารางที่ 11(ก) : การเปรียบเทียบค่าความถูกต้องของการจำแนกชุดข้อมูล Thaiprinted Character2	52
ตารางที่ 11(ข) : การเปรียบเทียบจำนวนครั้งของการจำแนกชุดข้อมูล Thaiprinted Character 2	52
ตารางที่ 12 : การเปรียบเทียบค่าความถูกต้องใน Polynomial kernel ของการจำแนก กรณีที่ให้ค่าความถูกต้องสูงสุดของแต่ละวิธี	54
ตารางที่ 13: การเปรียบเทียบค่าความถูกต้องใน RBF kernel ของการจำแนก กรณีที่ให้ค่าความถูกต้องสูงสุดของแต่ละวิธี.....	55
ตารางที่ 14: การเปรียบเทียบค่าจำนวนครั้งในการจำแนกกรณีที่ให้ค่าความถูกต้องสูงสุด ของแต่ละวิธี	55
ตารางที่ 15: การเปรียบเทียบค่าระดับความเชื่อมั่น (Confidence Level) ของค่าความต่างของ ค่าความถูกต้องระหว่างวิธีแตกครึ่งตามสารสนเทศกับวิธีอื่น	57