

การรวมตัวจำแนกด้วยการลงคะแนนร่วมกับกฎไกล์เคียง

โดย

นายอิทธิ รมณียางกูร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิทยาการคอมพิวเตอร์ ภาควิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2550

การรวมตัวจำแนกด้วยวิธีลงคะแนนร่วมกับกฎใกล้เคียง

โดย

อิทธิ รมณียางกูร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิทยาการคอมพิวเตอร์ ภาควิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2550

Ensemble with Neighbor Rules Voting

By

Itt Romneeyangkurn

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิทยาการคอมพิวเตอร์ ภาควิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2550

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

อิทธิ รมนียางกูร

เรื่อง

การรวมตัวจำแนกด้วยวิธีลงคะแนนร่วมกับกฎใกล้เคียง

ประธานกรรมการวิทยานิพนธ์

(ดร.ชลวิษ นัทธี)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์

(ดร.สุกรี สิ้นธุภิณฺญ)

กรรมการวิทยานิพนธ์

(ดร.วนิดา พุทธิวิทยา)

กรรมการวิทยานิพนธ์

(ดร.วนิดา พุทธิวิทยา)

คณบดี

(ดร.เด่นดวง ประดับสุวรรณ)

บทคัดย่อ

ได้มีการพิสูจน์มาแล้วว่า ผลที่ได้จากการรวมคำตอบจากต้นไม้ตัดสินใจหลายต้นให้ ความถูกต้องมากกว่าผลที่ได้จากต้นไม้ตัดสินใจต้นเดียว โดยหลายๆ วิธีของการผลิตต้นไม้ตัดสินใจหลายต้นเริ่มจากเตรียมชุดข้อมูลเรียนรู้สำหรับให้ต้นไม้ตัดสินใจแต่ละต้นเรียนรู้ ด้วยวิธีบูตสแตรปปีง โดยการสุ่มตัวอย่างจำนวนเท่ากับจำนวนตัวอย่างของชุดข้อมูลเรียนรู้ตั้งต้น ดังนั้นตัวอย่างบางตัวอย่างของข้อมูลเรียนรู้ตั้งต้น อาจจะปรากฏมากกว่าหนึ่งครั้งในขณะที่บาง ตัวอย่างอาจไม่ปรากฏเลยในชุดข้อมูลเรียนรู้ใหม่นี้ เพื่อให้ได้ข้อมูลหลายๆ ชุดและนำมาใช้ในการ ฝึกต้นไม้ตัดสินใจแต่ละต้น วิทยานิพนธ์ฉบับนี้นำเสนอความคิดการนำกฎใกล้เคียงที่ได้จากแต่ละ การจำแนกร่วมในการลงคะแนน แทนที่จะใช้กฎที่จำแนกตัวอย่างได้เท่านั้นเพราะจะเห็นได้ ว่าบูตสแตรปปีงเกิดจากการสุ่มข้อมูลจากข้อมูลเรียนรู้ตั้งต้นเดียวกัน ดังนั้นกฎที่ได้จากแต่ละการ จำแนกน่าจะมีความสัมพันธ์กัน จากการทดสอบพบว่าการที่เรานำกฎใกล้เคียงมาร่วมลงคะแนน ด้วยนั้นให้ผลที่มีความถูกต้องที่มากขึ้นกว่าการลงคะแนนโดยวิธีปกติ นอกจากนี้ยังสังเกตหาค่า ความใกล้เคียงที่น้อยที่สุดที่ทำให้ค่าความถูกต้องมากขึ้นอีกด้วย

Abstract

Ensembles of classifiers have been employed to improve accuracy over single classifier. Various methods sequentially bootstrap data set and invoke a base classifier on these different bootstraps. In this paper, we propose an idea based on the use of "similar rules" or "neighbor rules" in voting for a given test example, instead of using only the rule that matches with the test example. From our experimental results, we can conclude that our method achieves comparable accuracy and is significantly better than a regular majority vote. We also empirically derive the least of value of a similarity between rules that gives more accurate result.

กิตติกรรมประกาศ

กราบขอบพระคุณอย่างสูงที่สุดในความเมตตาและความสามารถของ ผศ.ดร. สุกรี สิ้นธุ
ภิญโญ อาจารย์ที่ปรึกษา

กราบขอบพระคุณอย่างสูงที่สุดในมารดา

กราบขอบพระคุณอย่างสูงที่สุดในครูบาอาจารย์ทุกท่านในชีวิตข้าพเจ้า

กราบขอบพระคุณอย่างสูงที่สุดในญาติพี่น้องทุกท่าน

กราบขอบพระคุณอย่างสูงที่สุดในเพื่อนๆ พี่ๆ น้องๆ

กราบขอบพระคุณอย่างสูงที่สุดในบริษัท บีบีคอนเทนต์แอนด์มีเดียเดียร์ จำกัด

ขอบคุณ พี่โต้ง, พี่เป้, พี่งค์, น้องเต๋า และ กิ๊ก

นายอิทธิ รมณียางกูร
มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2550

สารบัญ

	หน้า
บทคัดย่อ	(2)
กิตติกรรมประกาศ.....	(4)
สารบัญ	(5)
สารบัญตาราง.....	(8)
สารบัญภาพประกอบ	(9)
บทที่	
1. บทนำ	1
1.1 ความเป็นมาและความสำคัญของงานวิจัย	1
1.2 วัตถุประสงค์ของงานวิจัย	2
1.3 ขอบเขตของงานวิจัย.....	2
2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	4
2.1 ทฤษฎีที่เกี่ยวข้อง.....	4
2.1.1 ต้นไม้ตัดสินใจ.....	4
2.1.2 การเตรียมชุดตัวอย่างเรียนรู้แบบสแตปปีง	11
2.1.3 การรวมตัวจำแนก	12
2.2 งานวิจัยที่เกี่ยวข้อง	15
2.2.1 เค-เพื่อนบ้านที่ใกล้ที่สุด	15
2.2.2 กฎที่ใกล้ที่สุด	15
2.2.3 การลดจำนวนกฎ	15

บทที่	หน้า
3. วิธีการดำเนินงานวิจัย	26
3.1 โครงสร้างของระบบ	26
3.1.1 ส่วนการเตรียมการจำแนกหรือต้นไม้ตัดสินใจ.....	26
3.1.2 ส่วนของการเรียนรู้ของต้นไม้ตัดสินใจ	27
3.2 วิธีการวัดประสิทธิภาพ.....	30
4. ผลการทดลอง	31
4.1 ข้อมูลที่ใช้ในการทดลอง.....	31
4.2 ผลการทดลอง	33
4.2.1 ความถูกต้องระหว่างการลงคะแนนโดยกฎ ⁺ กับแบ็กกิง .	33
4.2.2 ความถูกต้องระหว่างการลงคะแนนโดยตัวอย่างเรียนรู้ ⁺ กับการ ลงคะแนนแบบทั่วไปโดยตัวอย่างเรียนรู้	35
4.3 วิเคราะห์ผลการทดลอง.....	35
5. สรุปผลการศึกษาและข้อเสนอแนะ.....	39
5.1 สรุปผลการศึกษาและวิจารณ์	39
5.2 ข้อเสนอแนะ.....	39
ภาคผนวก	หน้า
ก. คุณลักษณะชุดข้อมูลที่ใช้ในการทดสอบ.....	42
ก.1 ชุดข้อมูลเสียง (Audiology).....	42
ก.2 ชุดข้อมูลประชากรออสเตรเลีย (Australian).....	45
ก.3 ชุดข้อมูลสเกลสมดุลย์ (Balance-Scale)	46
ก.4 ชุดข้อมูลสะพาน (Bridges).....	47

ก.5 ชุดข้อมูลรถยนต์ (Car).....	48
ก.6 ชุดข้อมูลผิวหนังวิทยา (Dermatology)	48
ก.7 ชุดข้อมูลแฮรทท์ (Hayes-Roth)	50
ก.8 ชุดข้อมูลหัวใจ (Heart).....	50
ก.9 ชุดข้อมูลโรคตับอักเสบ (Hepatitis).....	51
ก.10 ชุดข้อมูลม้า (Horse-Colic).....	52
ก.11 ชุดข้อมูลคุณภาพชีวิตแรงงาน (Labor-Negotiations).....	54
ก.12 ชุดข้อมูลตับอักเสบ (Liver-Disorder).....	55
ก.13 ชุดข้อมูลนมถั่วเหลือง (Soybean)	55
ก.14 ชุดข้อมูลครูสอนภาษาอังกฤษ (TAE)	57
ก.15 ชุดข้อมูลสวนสัตว์ (Zoo).....	57
ข. โครงสร้างฐานข้อมูลระบบ.....	60
บรรณานุกรม	61
ประวัติการศึกษา.....	63

สารบัญตาราง

ตารางที่		หน้า
2.1	ตัวอย่างข้อมูลการเล่นเทนนิส	5
2.2	ชุดข้อมูลที่ได้ของวิธีแบ็กกิง	12
2.3	ชุดข้อมูลที่ได้ของวิธีบูสต์	14
4.1	ชุดข้อมูลที่ใช้ในการทดสอบงานวิจัย	32
4.2	การเปรียบเทียบถูกต้องระหว่างการลงคะแนนโดยกฎ ⁺ กับแบ็กกิง	34
4.3	การเปรียบเทียบถูกต้องระหว่างการลงคะแนนโดยตัวอย่างเรียนรู้ ⁺ กับการลงคะแนนแบบ ทั่วไปโดยตัวอย่างเรียนรู้.....	36
4.4	จำนวนข้อมูลการลงคะแนนโดยกฎ ⁺ เปรียบเทียบกับแบ็กกิง	37
4.5	จำนวนข้อมูลการลงคะแนนโดยตัวอย่าง ⁺ เปรียบเทียบกับการลงคะแนนแบบทั่วไปโดย ตัวอย่างเรียนรู้.....	37

สารบัญภาพประกอบ

ภาพที่		หน้า
2.1	ตัวอย่างต้นไม้ตัดสีเขียวของการทำนายเล่นเทนนิสหรือไม่.....	6
2.2	ความสัมพันธ์ของค่าเอนโทรปี.....	8
2.3	ผลลัพธ์ต้นไม้ตัดสีเขียวจากการหาค่าเพิ่มสารสนเทศขั้นแรก	11
2.4	การรวมตัวจำแนก.....	13
2.5	ขอบเขตค่าคุณลักษณะของสองกฎที่ไม่มีส่วนทับซ้อนกัน.....	21
2.6	ขอบเขตค่าคุณลักษณะของสองกฎที่มีส่วนทับซ้อนกัน	21
3.1	การเตรียมการจำแนกในแต่ละการจำแนกและผลลัพธ์.....	27
3.2	ส่วนของการรวมข้อมูลที่ได้ในแต่ละการจำแนก	28
3.3	อัลกอริทึมการลงคะแนนไปโดยกฎ ⁺	29
3.4	อัลกอริทึมการลงคะแนนโดยตัวอย่างเรียนรู้ ⁺	29
4.1	การแบ่งข้อมูลสำหรับการเรียนรู้และทดสอบ	31
4.2	การแบ่งข้อมูลจำนวน 10 ข้อมูลย่อย	33