ฟังก์ชันถ่วงน้ำหนักสำหรับอัลกอริทึมเนียเรสท์เนเบอร์

นายภาสกร ตั้งชนะชัยอนันต์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2548

ISBN 974-53-2873-1

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

I27677056

WEIGHT FUNCTIONS FOR NEAREST NEIGHBOR ALGORITHMS

Mr. Pasakorn Tangchanachaianan

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 2005

ISBN 974-53-2873-1

481627

หัวข้อวิทยานิพนธ์	ฟังก์ชันถ่วงน้ำหนักสำหรับอัลกอริทึมเนียเรสท์เนเบอร์
โดย	นายภาสกร ตั้งชนะชัยอนันต์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา	รองศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์

(ศาสตราจารย์ ดร. ดิเรก ลาวัณย์ศิริ)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(รองศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษา

(รองศาสตราจารย์ ดร. บุญเสริม กิจศิริกุล)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร. ญาใจ ลิ่มปิยะกรณ์)

..... กรรมการ

(อาจารย์ ดร. ชลวิช นัทธี)

ภาสกร ตั้งชนะชัยอนันต์ : ฟังก์ชันถ่วงน้ำหนักสำหรับอัลกอริทึมเนียเรสท์เนเบอร์. (WEIGHT FUNCTIONS FOR NEAREST NEIGHBOR ALGORITHMS) อ.ที่ปรึกษา: รศ. ดร. บุญเสริม กิจศิริกุล, 83 หน้า. ISBN 974-53-2873-1.

วิทยานิพนธ์ฉบับนี้มีวัตถุประสงค์ที่จะค้นหาฟังก์ชันถ่วงน้ำหนักที่ดีเพื่อใช้สำหรับวัดความแตกต่างระหว่างข้อมูลสองตัวใดๆ ในอัลกอริทึมเนียเรสท์เนเบอร์ โดยเน้นที่จะพัฒนาความแม่นยำของการจำแนกประเภทเป็นหลัก

กระบวนการที่ใช้เริ่มต้นโดยการให้นิยามเกี่ยวกับคุณสมบัติต่างๆ ของฟังก์ชันซึ่งเป็นที่ต้องการเพื่อหารูปแบบที่เหมาะสมของฟังก์ชันนี้ หลังจากนั้นรูปแบบของฟังก์ชันที่เป็นผลลัพธ์จากขั้นตอนที่ผ่านมา จะถูกนำมาทำการทดลองโดยการจำแนกประเภทกับชุดข้อมูลซึ่งจะถูกใช้เป็นตัววัดผล โดยมีวัตถุประสงค์เพื่อที่จะทดสอบประสิทธิภาพของฟังก์ชันที่ได้ ซึ่งในที่นี้ก็คือความแม่นยำในการจำแนกประเภทที่จะนำมาเปรียบเทียบกับฟังก์ชันถ่วงน้ำหนักแบบพื้นฐานที่ใช้กับอัลกอริทึมประเภทนี้ และเพื่อที่จะค้นหาค่าปรับแต่งที่เหมาะสมที่สุดของฟังก์ชันเหล่านั้น และผลการทดลองที่ได้ก็แสดงให้เห็นถึงประสิทธิภาพของฟังก์ชันนั้นซึ่งก็ถือได้ว่าพัฒนามีการพัฒนาขึ้นพอสมควรเมื่อเทียบกับฟังก์ชันถ่วงน้ำหนักแบบพื้นฐาน

ภาควิชา...วิศวกรรมคอมพิวเตอร์.....ลายมือชื่อนิสิต.....

สาขาวิชา...วิศวกรรมคอมพิวเตอร์.....ลายมือชื่ออาจารย์ที่ปรึกษา.....

ปีการศึกษา...2548.....

4770401221 : MAJOR COMPUTER ENGINEERING

KEY WORD: NEAREST NEIGHBOR / CLASSIFICATION / WEIGHT FUNCTION

PASAKORN TANGCHANACHAIANAN : WEIGHT FUCTIONS FOR NEAREST NEIGHBOR ALGORITHMS. THESIS ADVISOR : ASSOC. PROF. BOONSERM KIJSIRIKUL, Ph.D.. 83 pp. ISBN 974-53-2873-1.

The purpose of this thesis is to find some good weight functions for calculating the difference between any two data in nearest neighbor algorithms, especially focusing on the accuracy of classification.

The process starts by defining properties of weight functions needed for determining the forms of appropriate functions. After that, these forms, resulted from the previous step, are used in the experiments to classify the test data. The purpose of this is to evaluate the accuracy, compared to the traditional weight functions used in the nearest neighbor algorithms. We also study the most suitable value for the adjustable factors of these functions. The result shows the effectiveness of the proposed functions. It could be considered as a fair improvement compared to the traditional weight functions.

Department...Computer Engineering.....Student's signature..Pasakorn Tangchanachaianan
Field of study...Computer Engineering.....Advisor's signature.....
Academic year...2005.....

กิตติกรรมประกาศ

ขอขอบพระคุณ รองศาสตราจารย์.ดร. บุญเสริม กิจศิริกุล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ทำการให้ความช่วยเหลือและการสนับสนุนนานับประการ จนกระทั่งวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ขอขอบพระคุณ รองศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา ผู้ช่วยศาสตราจารย์ ดร. ญาใจ ลิ่มปิยะกรณ์ และ อาจารย์ ดร. ชลวิช นัทธี ที่ได้กรุณาสละเวลามาเป็นกรรมการสอบวิทยานิพนธ์ฉบับนี้

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ง
บทคัดย่อภาษาอังกฤษ ..... จ
กิตติกรรมประกาศ ..... ฉ
สารบัญ ..... ช
สารบัญตาราง ..... ฌ
สารบัญภาพ ..... ฎ

บทที่

1. บทนำ 1
    1.1 ความเป็นมาและความสำคัญของปัญหา 1
    1.2 วัตถุประสงค์ของการวิจัย 2
    1.3 ขอบเขตของวิทยานิพนธ์ 3
    1.4 ประโยชน์ที่คาดว่าจะได้รับ 3
    1.5 วิธีดำเนินการวิจัย 3
2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 4
    2.1 ทฤษฎีที่เกี่ยวข้อง 4
        2.1.1 แผนภาพโวโรนอย 4
        2.1.2 การตรวจสอบไขว้ 5
    2.2 งานวิจัยที่เกี่ยวข้อง 6
        2.2.1 อัลกอริทึมเคเนียเรสท์เนเบอร์ 6
        2.2.2 อัลกอริทึมเนียเรสท์เนเบอร์แบบถ่วงน้ำหนัก 7
        2.2.3 ฟังก์ชันระยะทาง 10
3. ฟังก์ชันถ่วงน้ำหนักสำหรับอัลกอริทึมเนียเรสท์เนเบอร์ 15
    3.1 คุณสมบัติเบื้องต้นของฟังก์ชันถ่วงน้ำหนักที่ต้องการ 15
    3.2 การทนทานต่อการสเกล 16
    3.3 ความมีประโยชน์ของข้อมูล 17
    3.4 การประมาณการกระจายตัวแบบสม่ำเสมอของข้อมูล 32

บทที่ หน้า

4. ผลการทดลอง 37
4.1 การประเมินผลการทดลอง 37
4.2 ผลที่ได้จากการทดลองและการวิเคราะห์ 39
5. สรุปผลการวิจัยและข้อเสนอแนะ 76
5.1 สรุปผลการวิจัย 76
5.2 ข้อเสนอแนะ 79
รายการอ้างอิง 81
ประวัติผู้เขียนวิทยานิพนธ์ 83

สารบัญตาราง

ตาราง หน้า

ตารางที่ 1 รายละเอียดเกี่ยวกับชุดข้อมูลจาก UCI Machine Learning Repository ที่นำมาใช้ในการทดสอบ 37

ตารางที่ 2 เปอร์เซ็นต์ความถูกต้องที่ได้จากอัลกอริทึมเคเนียเรสท์เนเบอร์ดั้งเดิม 39

ตารางที่ 3 เปอร์เซ็นต์ความถูกต้องที่ได้จากฟังก์ชันถ่วงน้ำหนัก $f(x)=\frac{1}{x^t}$ 40

ตารางที่ 4 เปอร์เซ็นต์ความถูกต้องที่ได้จากฟังก์ชันถ่วงน้ำหนักในรูปแบบที่ 1 41

ตารางที่ 5 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_1(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 3 ตัว 43

ตารางที่ 6 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_1(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 5 ตัว 44

ตารางที่ 7 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_1(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 10 ตัว 45

ตารางที่ 8 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_1(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 20 ตัว 46

ตารางที่ 9 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_1(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 40 ตัว 47

ตารางที่ 10 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_1(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 80 ตัว 48

ตารางที่ 11 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_1(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบเป็นข้อมูลสอนที่เหลือทุกตัว 49

ตารางที่ 12 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_2(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 3 ตัว 50

ตารางที่ 13 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_2(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 5 ตัว 51

ตารางที่ 14 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_2(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 10 ตัว 52

ตารางที่ 15 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_2(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 20 ตัว 53

ตารางที่ 16 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_2(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 40 ตัว 54
ตารางที่ 17 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_2(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 80 ตัว 55
ตารางที่ 18 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_2(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบเป็นข้อมูลสอนที่เหลือทุกตัว 56
ตารางที่ 19 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_3(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 3 ตัว 57
ตารางที่ 20 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_3(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 5 ตัว 58
ตารางที่ 21 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_3(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 10 ตัว 59
ตารางที่ 22 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_3(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 20 ตัว 60
ตารางที่ 23 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_3(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 40 ตัว 61
ตารางที่ 24 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_3(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบ 80 ตัว 62
ตารางที่ 25 เปอร์เซ็นต์ความถูกต้องจากการใช้ $h_3(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบเป็นข้อมูลสอนที่เหลือทุกตัว 63
ตารางที่ 26 สรุปเปอร์เซ็นต์ความถูกต้องจากการใช้ $h(d)$ กับฟังก์ชันรูปแบบที่ 1 เมื่ออาศัยข้อมูลโดยรอบจำนวนต่างๆ กัน k ตัว 64
ตารางที่ 27 เปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[k]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบ 3 ตัว 67
ตารางที่ 28 เปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[k]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบ 5 ตัว 68
ตารางที่ 29 เปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[k]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบ 10 ตัว 69
ตารางที่ 30 สรุปเปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[k]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบจำนวนต่างๆ กัน k ตัว 70

ตารางที่ 31 เปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[n]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบ 3 ตัว....71
ตารางที่ 32 เปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[n]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบ 5 ตัว....72
ตารางที่ 33 เปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[n]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบ 10 ตัว....73
ตารางที่ 34 สรุปเปอร์เซ็นต์ความถูกต้องจากฟังก์ชันรูปแบบที่ 2 เมื่อใช้ $\sqrt[n]{h_3(d)}$ โดยอาศัยข้อมูลโดยรอบจำนวนต่างๆ กัน k ตัว....74

สารบัญภาพ

ภาพประกอบ หน้า

รูปที่ 1 ตัวอย่างแผนภาพโวโรนอยใน 2 มิติ....4

รูปที่ 2 ผลของฟังก์ชันระยะทางแบบต่างๆ บนปริภูมิ 2 มิติ....13

รูปที่ 3 ภาพตัวอย่างของการมีประโยชน์ไม่มากกว่า....18

รูปที่ 4 ภาพตัวอย่างสนับสนุนแนวคิดในการใช้ค่าเฉลี่ยเรขาคณิต....33

รูปที่ 5 ผลลัพธ์ที่เป็นไปได้แต่ละชนิดของข้อมูลที่สร้างขึ้นเอง....38