วิธีการหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์

นายพงศกร ธีรรัศมี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2558

THE METHODOLOGY TO FIND APPROPRIATE K FOR K-NEAREST NEIGHBOR CLASSIFICATION WITH MEDICAL DATASETS

Pongsakorn Teerarassamee

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

Suranaree University of Technology

Academic Year 2015

วิธีการหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

_____

(รศ. ดร.กิตติศักดิ์ เกิดประสพ)
ประธานกรรมการ

_____

(รศ. ดร.นิตยา เกิดประสพ)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

_____

(ผศ. สมพันธุ์ ชาญศิลป์)
กรรมการ

มหาวิทยาลัยเทคโนโลยีสุรนารี

_____

(ศ. ดร.ชูกิจ ลิมปิจำนงค์)
รองอธิการบดีฝ่ายวิชาการและนวัตกรรม

_____

(รศ. ร.อ. ดร.กนต์ธร ชำนิประศาสน์)
คณบดีสำนักวิชาวิศวกรรมศาสตร์

พงศกร ธีรรัศมี : วิธีการหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์ (THE METHODOLOGY TO FIND APPROPRIATE K FOR K-NEAREST NEIGHBOR CLASSIFICATION WITH MEDICAL DATASETS)
อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.นิตยา เกิดประสพ, 72 หน้า

งานวิจัยนี้ได้ศึกษาปัญหาการจำแนกข้อมูลด้วยอัลกอริทึมเคเนียร์เรสเนเบอร์ ซึ่งในการจำแนกข้อมูลจำเป็นต้องกำหนดค่า เค ซึ่งหมายถึงจำนวนข้อมูลใกล้เคียง โดยค่า เค ที่กำหนดจะมีผลต่อประสิทธิภาพในการวิเคราะห์เพื่อจำแนกประเภทของข้อมูล หากกำหนดค่า เค ไม่เหมาะสมจะทำให้ค่าความแม่นจากการจำแนกประเภทต่ำกว่าที่ควรจะเป็น นอกจากนั้นหากกำหนดค่า เค มากเกินไปจะส่งผลให้การประมวลผลช้า และอาจทำให้ความแม่นลดลง การใช้วิธีปรับค่า เค ไปตามลำดับจนกระทั่งได้ค่าความแม่นที่สูง ไม่เป็นที่นิยมนักเนื่องจากจะใช้เวลาในการประมวลผลมากและอาจจะไม่ได้ค่าความแม่นสูงตามที่ต้องการ ดังนั้นงานวิจัยนี้ได้แนะนำวิธีการกำหนดค่า เค ที่เหมาะสมในการจำแนกข้อมูลทางการแพทย์ที่จะส่งผลให้การจำแนกข้อมูลมีค่าความแม่นสูง และเป็นการเพิ่มทางเลือกในการตัดสินใจให้กับผู้ใช้

การวิเคราะห์เพื่อให้ได้ซึ่งค่า เค ที่เหมาะสมสำหรับจำแนกประเภทข้อมูลด้วยอัลกอริทึมเคเนียร์เรสเนเบอร์ จำเป็นต้องรู้ถึงลักษณะของข้อมูล ซึ่งในงานวิจัยนี้จะมีการพิจารณาค่าทางสถิติเกี่ยวกับการกระจายของข้อมูล สำหรับใช้ในการวิเคราะห์เพื่อแนะนำการเลือกค่า เค เพื่อเป็นพารามิเตอร์สำคัญของอัลกอริทึมเคเนียร์เรสเนเบอร์

มหาวิทยาลัยเทคโนโลยีสุรนารี

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา_____

ปีการศึกษา 2558 ลายมือชื่ออาจารย์ที่ปรึกษา_____

PONGSAKORN TEERARASSAMEE : THE METHODOLOGY TO FIND APPROPRIATE K FOR K-NEAREST NEIGHBOR CLASSIFICATION WITH MEDICAL DATASETS. THESIS ADVISOR : ASSOC. PROF. NITTAYA KERDPRASOP, Ph.D., 72 PP.

K-NEAREST NEIGHBOR/ CLASSIFICATION/ MEDICAL DATA

In this research, we have studied the problem of data classification with the k-nearest neighbor (kNN) algorithm. On classifying data, the value of k, which means the number of closest data, has to be specified. The choice of k has great impact to the efficiency of data classification. The bad choice of k value results in a low classification accuracy. The too high value of k can also slow down the computation time and may decrease the accuracy. Sequentially adjusting k value until reaching the optimal one is not a practical method because it takes much processing time and the best accuracy cannot be guaranteed. Therefore, this research has suggested a way to configure the appropriate value of k based on data characteristics that can result in accurate classification. The proposed method can help users estimate appropriate value of k in a quick time.

On analyzing the k value suitable for kNN algorithm, it is necessary to know the data characteristics. In this research, we consider the distribution of data for analyzing and recommending the k value to be an important parameter for the kNN algorithm.

School of Computer Engineering Student's Signature_____

Academic Year 2015 Advisor's Signature_____

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอกราบขอบพระคุณบุคคล และกลุ่มบุคคลต่าง ๆ ที่ได้กรุณาให้คำปรึกษา แนะนำ ช่วยเหลืออย่างดียิ่ง ทั้งในด้านวิชาการ และด้านการดำเนินงานวิจัย ดังต่อไปนี้

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ และรองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ ที่ให้คำปรึกษาในการทำงานวิจัย การจัดการรูปแบบ และช่วยตรวจทานความถูกต้องของวิทยานิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.ชาญวิทย์ แก้วกสิ ผู้ช่วยศาสตราจารย์ ดร.คะชา ชาญศิลป์ ผู้ช่วยศาสตราจารย์ สมพันธุ์ ชาญศิลป์ และผู้ช่วยศาสตราจารย์ ดร.ปรเมศวร์ ห่อแก้ว อาจารย์ประจำสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี

คุณกัลญา พับโพธิ์ เลขานุการสาขาวิชาวิศวกรรมคอมพิวเตอร์ ที่ให้ความช่วยเหลือในการประสานงานด้านเอกสารระหว่างศึกษา

คุณภาสพิชญ์ ชูใจ คุณไพชยนต์ คงไชย คุณกีระชาติ สุขสุทธิ์ คุณกิตติพงศ์ ชมบุญ คุณนันทวุฒิ คะอังกุ คุณศักย์ เพิ่มหรรษา คุณรติพร จันทร์กลั่น และนักศึกษาบัณฑิตสาขาวิชาวิศวกรรมคอมพิวเตอร์ ทุกท่านที่ให้คำปรึกษาและช่วยเหลือด้วยดีมาโดยตลอด

นอกจากนี้ขอขอบคุณครู อาจารย์ทั้งในอดีตและปัจจุบันที่ให้ความรู้แก่ผู้วิจัยจนประสบความสำเร็จในชีวิต

ท้ายที่สุดที่จะลืมไม่ได้ ขอกราบขอบพระคุณ คุณนิตยา ธีรรัศมี มารดาของข้าพเจ้าที่ให้กำเนิด อบรม เลี้ยงดูด้วยความรัก และส่งเสริมการศึกษาเป็นอย่างดีโดยตลอด ทำให้ผู้วิจัยมีความรู้ความสามารถ มีจิตใจที่เข้มแข็ง รวมทั้งเป็นกำลังใจที่ยิ่งใหญ่แก่ผู้วิจัย จนทำให้ผู้วิจัยประสบความสำเร็จในชีวิตเรื่อยมา

พงศกร ธีรรัศมี

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

มหาวิทยาลัยเทคโนโลยีสุรนารี

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

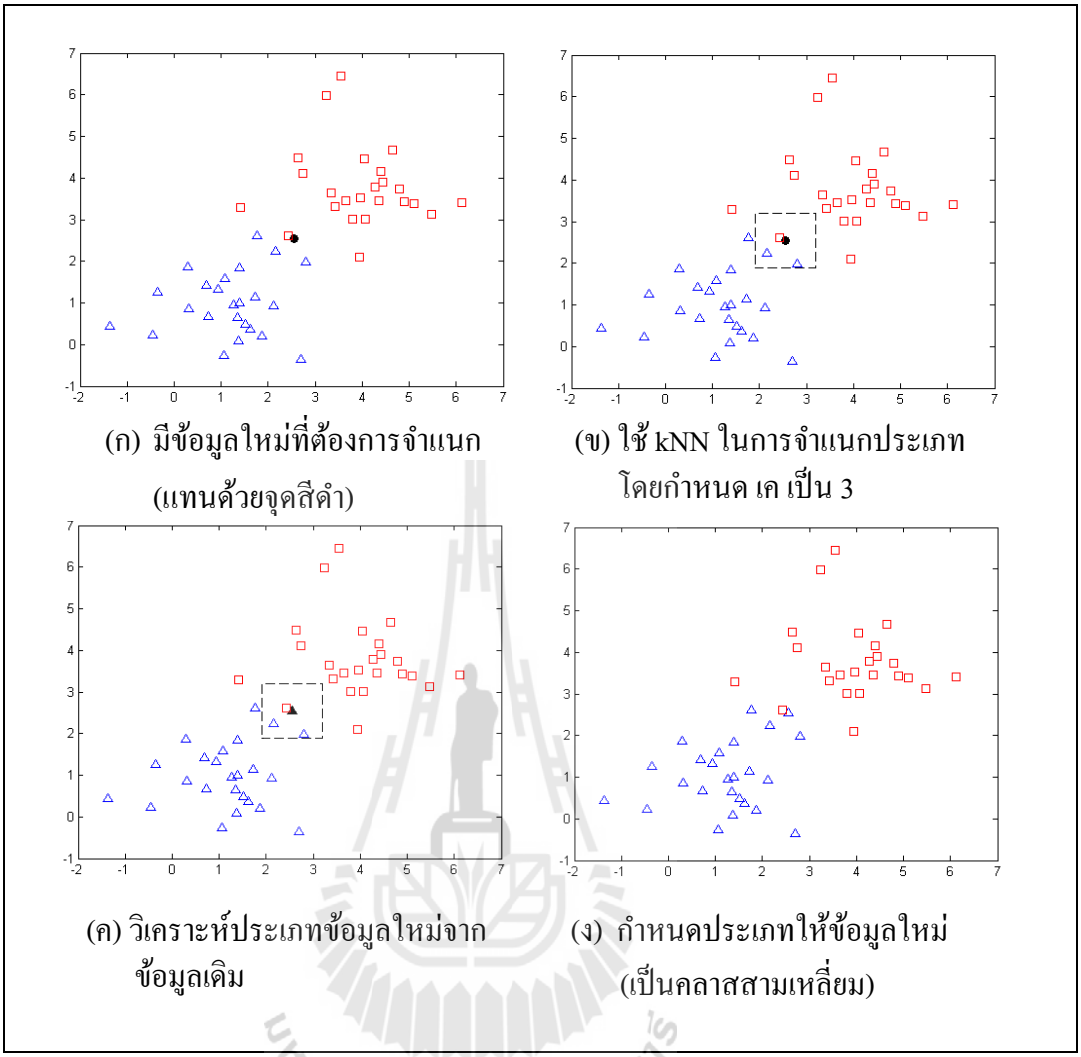
มหาวิทยาลัยเทคโนโลยีสุรนารี

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหางานวิจัย

เคเนียร์เรสเนเบอร์ หรือ k-Nearest Neighbor (kNN) เป็นวิธีการจำแนกประเภทข้อมูล (Data Classification) วิธีการหนึ่ง โดยจัดเป็นวิธีการจำแนกประเภทข้อมูลแบบมีผู้ฝึกสอน (Supervised Learning) หรือการที่ทราบคำตอบของข้อมูลอยู่ก่อนแล้ว จากนั้นใช้โมเดลในการจำแนกประเภทข้อมูลจากข้อมูลฝึกที่ทราบคำตอบ วิธีการจำแนกของเคเนียร์เรสเนเบอร์จะใช้วิธีการวิเคราะห์จากข้อมูลที่ใกล้เคียงที่สุดจำนวน เค ตัว กับข้อมูลที่ต้องการจำแนกประเภทของข้อมูลหรือต้องการทำนายคลาสของข้อมูลใหม่ โดยจะทำนายตามคลาสส่วนใหญ่ของข้อมูลฝึก เค ตัว ซึ่งเทคนิคดังกล่าวนิยมใช้กันอย่างแพร่หลายในหลายด้าน เช่น ด้านการแพทย์ (Hu and Shao, 2012) ด้านอุทกวิทยาและอุตุนิยมวิทยา (Lee and Ouarda, 2011) เป็นต้น

การกำหนดค่า เค ให้กับเคเนียร์เรสเนเบอร์จึงเป็นการกำหนดขอบเขตหรือบริเวณในการวิเคราะห์ข้อมูล โดยการกำหนดค่า เค นี้ มีผลเกี่ยวกับการใช้ทรัพยากรในการวิเคราะห์ข้อมูล ซึ่งถ้าหากมีการกำหนดค่า เค ที่สูงจะส่งผลให้มีการประมวลผลที่ใช้เวลานานขึ้น จากรูปที่ 1.1 เมื่อมีข้อมูลใหม่ที่ต้องการจำแนกประเภทโดยใช้เคเนียร์เรสเนเบอร์ขั้นตอนแรกจำเป็นต้องมีการกำหนดค่า เค ซึ่งในกรณีนี้กำหนดค่า เค เป็น 3 (รูปที่ 1.1 ก) ดังนั้นข้อมูลที่นำมาวิเคราะห์ในการจำแนกประเภทจะมี 3 ข้อมูล ซึ่งเป็นขอบเขตในการจำแนกประเภทของข้อมูลใหม่ ตามค่า เค ที่ถูกกำหนด (รูปที่ 1.1 ข) เมื่อได้ขอบเขตในการวิเคราะห์ข้อมูลแล้วเคเนียร์เรสเนเบอร์จะจำแนกข้อมูลใหม่ตามประเภทของข้อมูลเดิมที่มี โดยในตัวอย่างนี้จะเห็นว่าข้อมูลใหม่ถูกจำแนกให้เป็นข้อมูลประเภทสามเหลี่ยม (รูปที่1.1 ค) เพราะในขอบเขตที่ เค เป็น 3 มีข้อมูล สามเหลี่ยม 2 ข้อมูล และสี่เหลี่ยม 1 ข้อมูลเคเนียร์เรสเนเบอร์จึงจำแนกให้เป็นไปตามประเภทของข้อมูลเดิมที่มีจำนวนมากกว่า เมื่อจำแนกประเภทเสร็จสิ้นจะได้ข้อมูลใหม่เป็นประเภทสามเหลี่ยม (รูปที่ 1.1 ง)



(ก) มีข้อมูลใหม่ที่ต้องการจำแนก (แทนด้วยจุดสีดำ)

(ข) ใช้ kNN ในการจำแนกประเภท โดยกำหนด เค เป็น 3

(ค) วิเคราะห์ประเภทข้อมูลใหม่จากข้อมูลเดิม

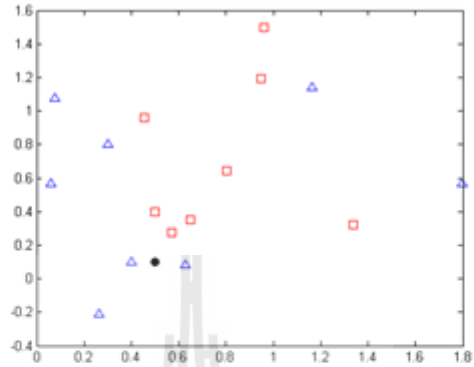
(ง) กำหนดประเภทให้ข้อมูลใหม่ (เป็นคลาสสามเหลี่ยม)

รูปที่ 1.1 การทำงานของ k-Nearest Neighbor

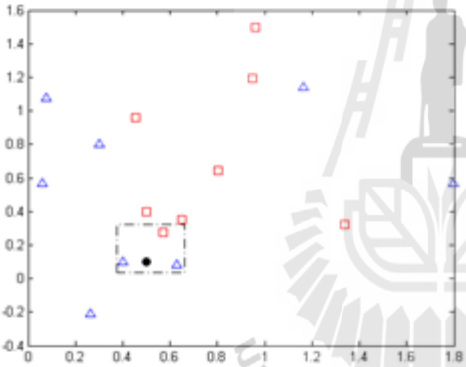
การกำหนดค่า เค ของเคเนียร์เรสเนเบอร์คือการกำหนดว่าจะวิเคราะห์ข้อมูลที่ใกล้กับข้อมูลที่ต้องการจำแนกที่สุดกี่ข้อมูล ซึ่งการเลือกจำนวนของค่า เค ที่แตกต่างกันนั้นจะให้ค่าความแม่น (Accuracy) ที่แตกต่างกันออกไป โดยการที่จะเลือกค่า เค ในการจำแนกข้อมูลให้ได้ค่าความแม่นสูงนั้น อาจจะต้องเสียเวลาในการปรับค่า เค ทีละค่า และเมื่อได้ค่าความแม่นที่ระดับหนึ่งแล้ว จะไม่สามารถทราบได้เลยว่าค่า เค ถัดไปจะให้ค่าความแม่นมากกว่าหรือไม่ และการเพิ่มค่า เค ไปในลักษณะลองผิดลองถูกจะส่งผลให้ การประมวลจะล่าช้า

ปัญหาของการกำหนดค่า เค ที่ไม่เหมาะสมจะส่งผลให้ค่าความแม่นต่างกัน ดังรูปที่ 1.2 (ก ถึง ค) ซึ่งเป็นการกำหนดค่า เค เป็น 3 เมื่อพิจารณาประเภทข้อมูลให้แก่ข้อมูลใหม่แล้ว จะเห็นว่าข้อมูลใหม่เป็นประเภทสามเหลี่ยม แต่หากกำหนดค่า เค เป็น 5 การพิจารณาประเภทของข้อมูลใหม่

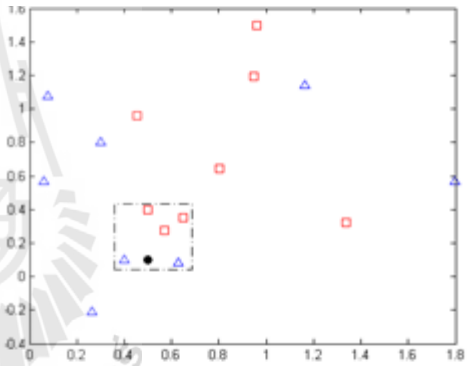
จะเป็นประเภทสี่เหลี่ยม โดยปกติแล้วการกำหนดค่า เค จะมีผลทำให้การจำแนกประเภทข้อมูลแตกต่างกันออกไปตามค่า เค ที่กำหนด ดังตัวอย่างข้างต้น ซึ่งหากต้องการจำแนกประเภทที่ให้ค่าความแม่นสูง โดยพิจารณาจากค่า เค จึงจำเป็นต้องมีวิธีในการเลือกค่า เค ที่เหมาะสม



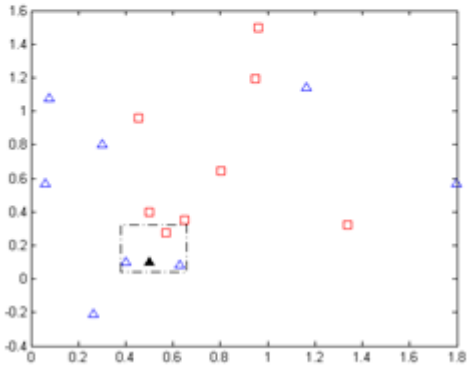
(ก) มีข้อมูลใหม่ที่ต้องการจำแนก (แทนด้วยจุดสีดำ)



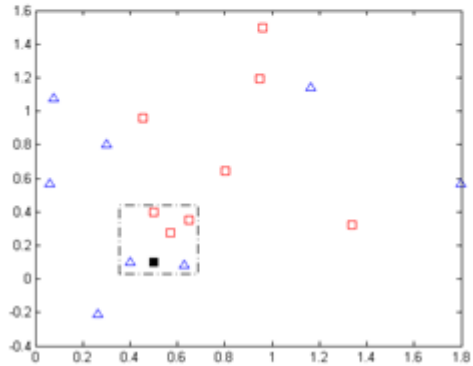
(ข) ใช้ kNN ในการจำแนกประเภทโดยกำหนด เค เป็น 3



(ง) ใช้ kNN ในการจำแนกประเภทโดยกำหนด เค เป็น 5



(ค) จำแนกข้อมูลใหม่เป็นคลาสสามเหลี่ยม



(จ) จำแนกข้อมูลใหม่เป็นคลาสสี่เหลี่ยม

รูปที่ 1.2 ปัญหาของการกำหนดค่า เค ที่ไม่เหมาะสมของ k-Nearest Neighbor

ในงานวิจัยนี้จึงได้นำเสนอเกี่ยวกับการแนะนำในการกำหนดค่า เค ที่เหมาะสมของเคเนียร์เรสเนเบอร์ ซึ่งใช้กับข้อมูลทางการแพทย์ เช่น ข้อมูลเกี่ยวกับผู้ป่วยโรคหอบหืดในระดับที่แตกต่างกัน ข้อมูลผู้ป่วยโรคหัวใจ และข้อมูลโรคมะเร็งเต้านม เป็นต้น นอกจากนี้มาตรวัดระยะทาง (Distance) ที่แตกต่างกันก็ให้ผลลัพธ์ที่แตกต่างกัน ค่า เค ที่เหมาะสมนอกจากช่วยให้ค่าความแม่นที่สูงในการจำแนกประเภทข้อมูลแล้วยังช่วยลดเวลาในการประมวลผลอีกด้วย ทำให้ผู้ที่ต้องการใช้งานเคเนียร์เรสเนเบอร์ในการจำแนกประเภทข้อมูลสามารถนำวิธีการจากงานวิจัยนี้ไปใช้ในการกำหนดค่า เค ได้ทันที

1.2 วัตถุประสงค์ของงานวิจัย

ผู้วิจัยได้ตั้งวัตถุประสงค์ในงานวิจัยไว้ คือการแนะนำเกี่ยวกับค่า เค ที่เหมาะสมในการจำแนกข้อมูลให้มีค่าความแม่นที่สูง โดยเสนอแนวทางการกำหนดค่า เค กับมาตรวัดระยะทางที่ควรจะใช้ในการจำแนกข้อมูล

1.3 ขอบเขตของงานวิจัย

1) ข้อมูลที่ใช้ในการทดลองประกอบด้วยข้อมูลดังต่อไปนี้
   - ข้อมูลผู้ป่วยที่เป็นโรคหอบหืดในระดับความรุนแรงที่ต่างกัน (ข้อมูลได้จากโรงพยาบาลมหาราช จังหวัดนครราชสีมา รวบรวมข้อมูล ณ วันที่ 8 พฤศจิกายน พ.ศ. 2557)
   - ข้อมูลผู้ป่วยเป็นโรคหัวใจ (https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart))
   - ข้อมูลผู้ป่วยเป็นโรคมะเร็งเต้านม ในวิสคอนซิน (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original))
   - ข้อมูลผู้ป่วยโรคไทรอยด์ (https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease)
   - ข้อมูลผู้ป่วยโรคเบาหวาน (https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes)
   - ข้อมูลสังเคราะห์
2) ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลที่มีลักษณะเป็นตัวเลข (Numeric) และคลาสเป้าหมาย (Target Class) เป็นข้อความ (Nominal)
3) ข้อมูลที่ใช้เป็นข้อมูล 2 คลาสเป้าหมาย หรือ 3 คลาสเป้าหมายเท่านั้น

4) งานวิจัยนี้เลือกใช้ภาษา MATLAB และภาษา R ในการพัฒนาและทดสอบอัลกอริทึมเคเนียร์เรสเนเบอร์
5) งานวิจัยนี้จะนำเสนอเกี่ยวกับการแนะนำในการกำหนดค่า เค ของเคเนียร์เรสเนเบอร์และมาตรวัดระยะทาง โดยใช้ข้อมูลทางการแพทย์ในการทดลอง

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่เกิดขึ้นจากงานวิจัยนี้ ประกอบด้วย

1) เป็นทางเลือกให้แก่ผู้ใช้เพื่อช่วยในการตัดสินใจ เกี่ยวกับการกำหนดค่าพารามิเตอร์ของเคเนียร์เรสเนเบอร์ ซึ่งประกอบด้วย มาตรวัดระยะทาง และค่า เค
2) เมื่อผู้ใช้เลือกใช้งานวิจัยนี้เพื่อช่วยกำหนดพารามิเตอร์ของเคเนียร์เรสเนเบอร์แล้วจะทำให้ผู้ใช้ไม่ต้องเสียเวลาในการเลือกค่า เค และมาตรวัดระยะทาง เพื่อได้ผลลัพท์ที่มีค่าความแม่นสูง

มหาวิทยาลัยเทคโนโลยีสุรนารี

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดของอัลกอริทึมเคเนียร์เรสเนเบอร์ การเลือกค่า เค ที่เหมาะสม มาตรวัดระยะทาง ประสิทธิภาพของในการจำแนกประเภทข้อมูล และงานวิจัยที่เกี่ยวข้อง

2.1 เคเนียร์เรสเนเบอร์

เคเนียร์เรสเนเบอร์ (k-Nearest Neighbor) เป็นวิธีในการจำแนกประเภทของข้อมูลวิธีการหนึ่งที่เป็นที่นิยมในการจำแนกประเภทข้อมูล เพราะลักษณะพิเศษของเคเนียร์เรสเนเบอร์คือ Lazy-Learning ซึ่งเหตุที่ถูกเรียกว่าเป็น Lazy-Learning นั้นมาจากรูปแบบในการจำแนกประเภทข้อมูลของเคเนียร์เรสเนเบอร์โดยไม่มีการสร้างโมเดลสำหรับจำแนกประเภทข้อมูลเตรียมไว้ล่วงหน้าเมื่อมีข้อมูลใหม่ที่ต้องการจำแนกประเภท เพียงนำมาเทียบกับข้อมูลเดิมและดูถึงความคล้ายคลึงกันของข้อมูลใหม่และข้อมูลเดิมที่มี ก็สามารถจำแนกประเภทของข้อมูลใหม่ได้ โดยให้เป็นประเภทเดียวกับข้อมูลเดิมที่อยู่ใกล้เคียง (Hulett et al., 2012)

ในทางการแพทย์นิยมใช้การจำแนกข้อมูลด้วยเคเนียร์เรสเนเบอร์เพราะการศึกษาเพื่อวินิจฉัยผลทางการแพทย์นิยมใช้เป็นลักษณะ Case Based Reasoning ซึ่งมีลักษณะเช่นเดียวกับเคเนียร์เรสเนเบอร์ที่หาความรู้ใหม่ จากข้อมูลเดิมที่มี ดังการวินิจฉัยที่ดูลักษณะอาการของผู้ป่วยเดิมเพื่อวินิจฉัยผู้ป่วยใหม่ ก่อนการประมวลผลข้อมูลเพื่อการจำแนกประเภทของเคเนียร์เรสเนเบอร์จะต้องมีการกำหนดค่า เค ซึ่งค่า เค หมายถึงจำนวนของข้อมูลเดิมที่ใกล้เคียงกับข้อมูลที่ต้องการจำแนกประเภทการเลือกค่า เค ที่เหมาะสม จะพิจารณาจากการวิเคราะห์ลักษณะของคลาสและข้อมูลเดิมที่มี ยกตัวอย่างเช่น ข้อมูลเดิมมีลักษณะเฉพาะเจาะจงและมีจำนวนคลาสน้อย การจำแนกเพื่อให้ได้ค่าความแม่นสูง อาจไม่จำเป็นต้องเลือกค่า เค ที่สูง ในทางกลับกันหากเลือกค่า เค เป็น 1 ค่าความแม่นควรจะสูงด้วย เพราะข้อมูลมีลักษณะเฉพาะและจำนวนคลาสน้อยทำให้แยกประเภทได้ง่าย

จากที่กล่าวถึงลักษณะการทำงานของเคเนียร์เรสเนเบอร์ข้างต้น ซึ่งลักษณะการทำงานจะเป็นการพิจารณาจากข้อมูลที่มีอยู่ก่อน และเมื่อต้องการระบุถึงข้อมูลใหม่ที่ต้องการจำแนกประเภทว่า

เป็นประเภทข้อมูลแบบใด เคเนียร์เรสเนเบอร์จะพิจารณาลักษณะที่ใกล้เคียงที่สุดกับข้อมูลเดิม และระบุว่าข้อมูลใหม่ที่ต้องการระบุประเภทนั้นเป็นประเภทใด ซึ่งรหัสเทียม (Pseudo Code) ของเคเนียร์เรสเนเบอร์ (Bunheang et al., 2014) มีลักษณะดังรูปที่ 2.1

k-Nearest Neighbor

Determine (k, distance)

Classify ($\boldsymbol{X}$, $\boldsymbol{Y}$, x) // $\boldsymbol{X}$ is training data, $\boldsymbol{Y}$ is labels of $\boldsymbol{X}$, x is unknown sample

for i = 1 **to** n **do** // n is all the training data

    Compute distance d($\boldsymbol{X}_i$,x)

end for

Compute set I containing indices for the k smallest distance d($\boldsymbol{X}_i$,x).

return majority label for {$\boldsymbol{Y}$ where i ∈ I}

รูปที่ 2.1 รหัสเทียมของเคเนียร์เรสเนเบอร์

จากรูปที่ 2.1 เป็นรหัสเทียมของเคเนียร์เรสเนเบอร์ ซึ่งสามารถอธิบายเป็นขั้นตอนง่าย ๆ ได้ 5 ขั้นตอนดังนี้

1. กำหนดค่า เค และมาตรวัดระยะทางให้กับเคเนียร์เรสเนเบอร์
2. คำนวณระยะทางระหว่างข้อมูลใหม่ที่ต้องการจำแนกกับข้อมูลเดิมที่มีทั้งหมด
3. เรียงลำดับระยะทาง และกำหนดเพื่อนบ้านที่ใกล้ที่สุดตามค่า เค
4. รวบรวมคลาสเป้าหมายของเพื่อนบ้าน
5. กำหนดคลาสให้กับข้อมูลใหม่โดยพิจารณาจากประเภทคลาสเป้าหมายของเพื่อนบ้านว่าเป็นประเภทใดมากที่สุด การกำหนดคลาสให้ข้อมูลใหม่ก็จะเป็นคลาสนั้น

2.2 มาตรวัดประสิทธิภาพในการจำแนกประเภทข้อมูล

โดยปกติแล้วการวัดประสิทธิภาพการจำแนกประเภทของข้อมูลสามารถวัดได้หลายวิธี เช่น Precision, Recall เป็นต้น แต่ในวิทยานิพนธ์นี้เลือกที่จะใช้ค่าความแม่น (Accuracy) เนื่องจากงานวิจัยนี้มีวัตถุประสงค์ในการแนะนำการกำหนดค่า เค เพื่อให้ได้ซึ่งค่าความแม่นที่สูง ซึ่งสามารถหาค่าความแม่นได้จาก Confusion Matrix โดยประกอบด้วยผลจากการทำนาย เปรียบเทียบกับ

ประเภทที่แท้จริงของข้อมูล โดยที่ Confusion Matrix มีลักษณะเป็นตาราง หากข้อมูลที่ต้องการจำแนกมี 2 ประเภท คือ Positive และ Negative ตาราง Confusion Matrix จะมีลักษณะดังรูปที่ 2.2

		Predicted Class	
		Positive	*Negative*
Actual Class	*Positive*	TP	FN
	Negative	FP	TN

รูปที่ 2.2 Confusion Matrix

จากรูปที่ 2.2 เมื่อมีการทำนาย 2 ประเภท ผลการทำนายทั้งหมดที่เป็นไปได้จะมี 4 ชนิด ดังนี้

- TP คือ ประเภทที่แท้จริงคือ Positive และการทำนายเป็น Positive
- TN คือ ประเภทที่แท้จริงคือ Negative และการทำนายเป็น Negative
- FP คือ ประเภทที่แท้จริงคือ Negative และการทำนายเป็น Positive
- FN คือ ประเภทที่แท้จริงคือ Positive และการทำนายเป็น Negative

ค่าความแม่นสามารถคำนวณได้จากสมการที่ 2-1 (Peterson, 2009) หรือหากพิจารณาจาก Confusion Matrix รูปแบบการคำนวณจะเป็นไปตามสมการที่ 2-2 โดยที่ทั้งสองสมการข้างต้นให้ผลลัพธ์ที่ไม่แตกต่างกัน

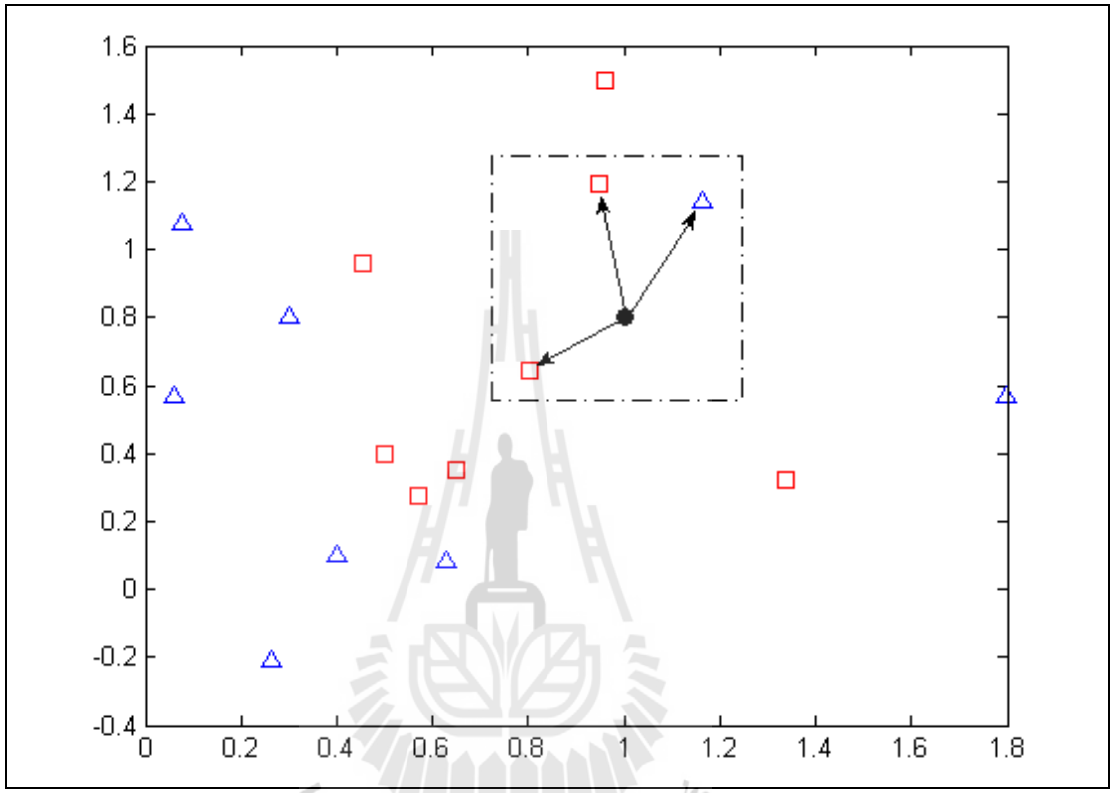
$$\text{Accuracy} = \frac{\text{จำนวนข้อมูลที่ทำนายคลาสได้ถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมด}} \quad (2\text{-}1)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2\text{-}2)$$

2.3 มาตรวัดระยะทาง

นอกจากมาตรวัดประสิทธิภาพการจำแนกประเภทข้อมูลที่ใช้ในการวัดค่าความแม่นในการจำแนกแล้ว มาตรวัดระยะห่างระหว่างข้อมูลยังมีผลในการวิเคราะห์เพื่อจำแนกประเภทอีกด้วย ตัวอย่างเช่น หากกำหนดค่า เค เป็น 3 แล้วเลือกใช้มาตรวัดระยะทางที่แตกต่างกันก็จะทำให้ค่าความแม่นที่ได้แตกต่างกันอีกด้วย ซึ่งเคเนียร์เรสเนเบอร์สามารถใช้มาตรวัดระยะทางต่าง ๆ ได้หลากหลายในการวิเคราะห์เพื่อจำแนกประเภทข้อมูลใหม่ที่ต้องการจำแนกโดยเทียบเคียงกับข้อมูล

เดิมที่อยู่บริเวณรอบ ๆ ตามค่า เค ที่กำหนด ดังรูป 2.2 หากกำหนดค่า เค เป็น 3 ลูกศรทั้ง 3 เส้นจะหมายถึงระยะทางระหว่างข้อมูลเดิมที่ทราบคลาสกับข้อมูลใหม่ที่ต้องการจำแนกประเภท (ข้อมูลใหม่แทนด้วยจุดสีดำ)



รูปที่ 2.3 ระยะทางจากข้อมูลใหม่ไปยังข้อมูลเดิม

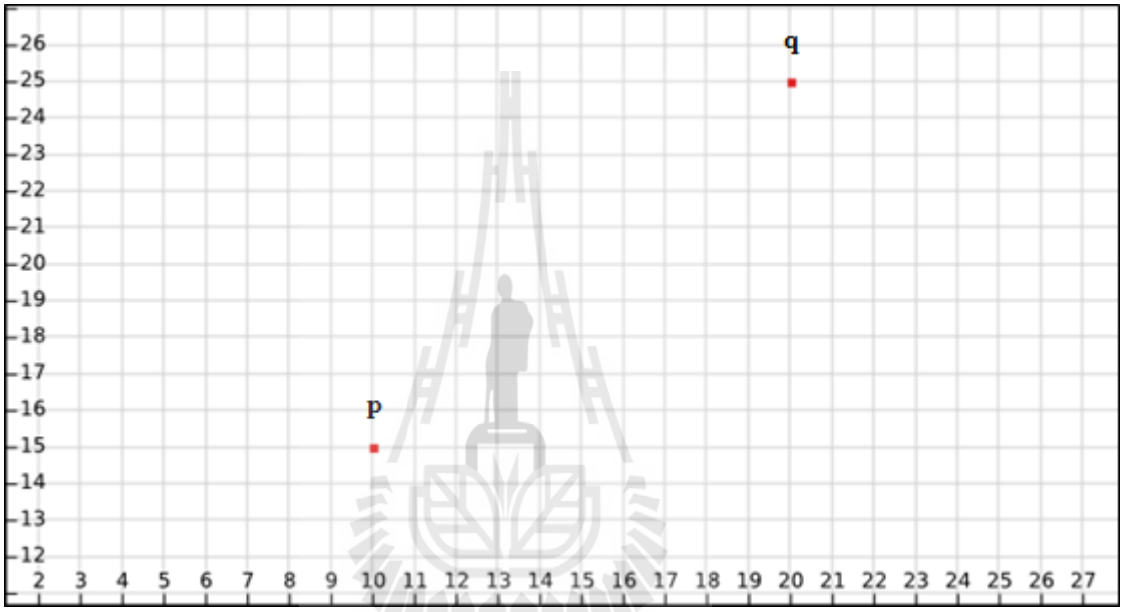
มาตรวัดระยะทางมีอยู่หลากหลายมาตรวัด แต่ในวิทยานิพนธ์นี้เลือกใช้มาตรวัดระยะทางที่ได้รับความนิยม ตามงานวิจัยของ S. A. Medjahed และคณะ (2013) ซึ่งประกอบด้วย มาตรวัดระยะทางดังต่อไปนี้

2.3.1 มาตรวัด Euclidean Distance

Euclidean Distance เป็นมาตรวัดระยะพื้นฐานใช้สำหรับหาระยะทางระหว่างจุดสองจุด เป็นที่นิยมใช้ในงานประเภทต่าง ๆ เป็นอย่างมาก เพราะง่ายต่อความเข้าใจ และลักษณะการคำนวณที่คล้ายกับทฤษฎีบทพีทาโกรัส (Deza and Deza, 2009) ซึ่งสามารถคำนวณได้ตามสมการที่

$$d_1(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2\text{-}3)$$

$d_1(p,q)$	คือระยะทางจากจุด p ไปยังจุด q วัดในแบบ Euclidean
p	คือจุดใด ๆ
q	คือจุดใด ๆ
n	คือจำนวนมิติของข้อมูล



รูปที่ 2.4 ตัวอย่างข้อมูลที่ใช้ในการคำนวณมาตรวัดระยะทาง

จากรูปที่ 2.4 แสดงตัวอย่างข้อมูลที่ใช้ในการคำนวณมาตรวัดระยะทางโดย กำหนดให้มีจุด p อยู่ที่พิกัดในแนวแกน x และ y เป็น (10,15) และจุดอยู่ที่พิกัด q (20,25) การคำนวณระยะห่างระหว่างจุด p และ q ด้วยวิธีการ Euclidean Distance สามารถแสดงขั้นตอนการคำนวณได้ดังต่อไปนี้

$$d_1(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

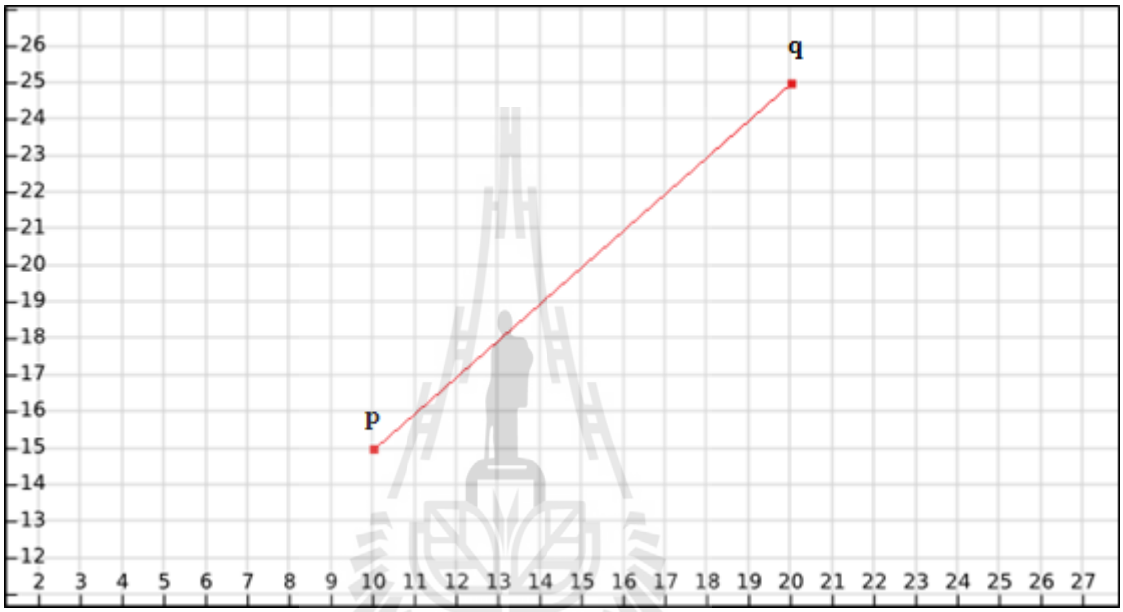
$$= \sqrt{(10-20)^2 + (15-25)^2}$$

$$= \sqrt{100+100}$$

$$= 10\sqrt{2}$$

$$= 14.142$$

จากตัวอย่างการวัดระยะทางโดยใช้มาตรวัด Euclidean Distance จะได้ผลลัพธ์ดังรูปที่ 2.5



รูปที่ 2.5 ตัวอย่างลักษณะการวัดระยะทางโดยใช้ Euclidean Distance

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.3.2 มาตรวัด City Block Distance

City Block Distance หรือเรียกอีกชื่อหนึ่งว่า Manhattan Distance เป็นมาตรวัดระยะทางที่มีลักษณะการวัดแบบทางเดินรถโดยระยะทางเป็นผลรวมด้านประกอบของ Euclidean Distance (Krause, 1987) ซึ่งสามารถคำนวณได้ตามสมการที่ 2-4

$$d_2(p,q) = \sum_{i=1}^{n} |p_i - q_i| \qquad (2\text{-}4)$$

$d_2(p,q)$	คือระยะทางจากจุด p ไปยังจุด q วัดในแบบ City Block
p	คือจุดใด ๆ
q	คือจุดใด ๆ
n	คือจำนวนจำนวนมิติของข้อมูล

จากรูปที่ 2.4 แสดงตัวอย่างการใช้มาตรวัด City Block Distance ในการวัดระยะทางจะคำนวณได้ดังนี้

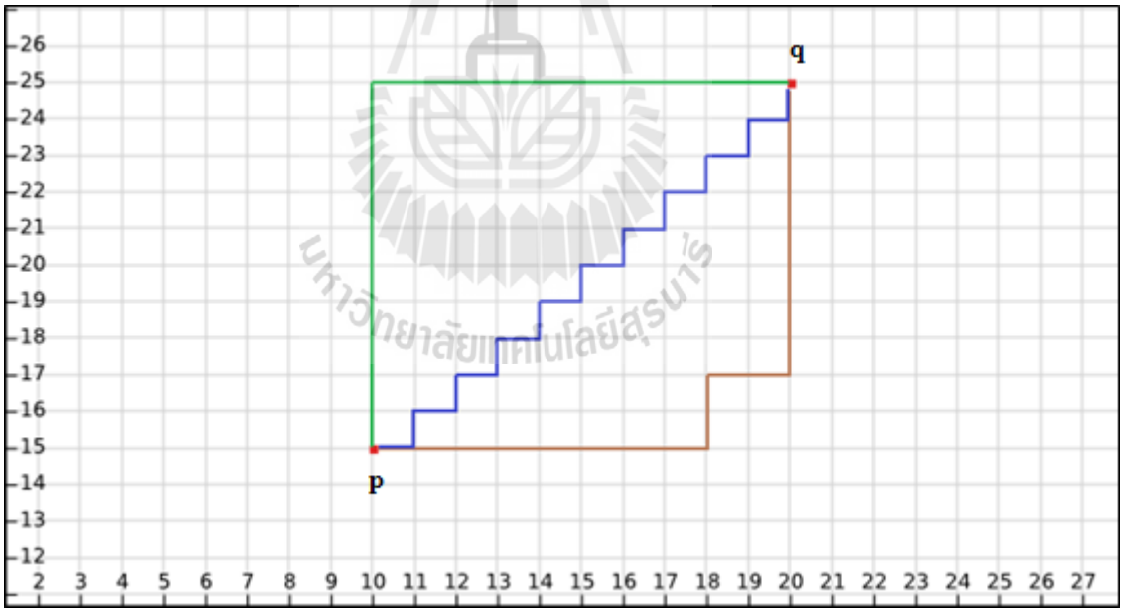
$$d_2(p,q) = \sum_{i=1}^{n} |p_i - q_i|$$

$$= |10 - 20| + |15 - 25|$$

$$= 10 + 10$$

$$= 20$$

เนื่องจากมาตรวัด City Block Distance มีลักษณะการวัดแบบการเดินรถ ซึ่งการเดินรถสามารถเลือกเส้นทางเดินได้หลายเส้นทาง และจากตัวอย่างการคำนวณก็สามารถวัดจากจุดหนึ่งไปอีกจุดหนึ่งได้หลายเส้นทางเช่นกัน ซึ่งมีลักษณะเส้นทางดังรูปที่ 2.6 โดยแสดงตัวอย่างเป็นสามเส้นทาง ดังเส้นสีเขียว เส้นสีน้ำเงิน และเส้นสีแดง



รูปที่ 2.6 ตัวอย่างลักษณะการวัดระยะทางโดยใช้ City Block Distance

2.3.3 มาตรวัด Cosine Distance

Cosine Distance (Tan et al., 2005) เกิดจากการคำนวณ Cosine Similarity สามารถหาค่าได้โดยการกระทำ Dot Product ของ 2 เวกเตอร์ และหากนำ 1 หักลบด้วยค่า Cosine Similarity ก็จะได้ Cosine Distance แสดงดังสมการที่ 2-9

$$d_3(p,q) = 1 - \frac{\sum_{i=1}^{n} p_i * q_i}{\sqrt{\sum_{i=1}^{n}(p_i)^2} * \sqrt{\sum_{i=1}^{n}(q_i)^2}} \quad (2\text{-}9)$$

$d_3(p,q)$	คือระยะทางจากจุด p ไปยังจุด q วัดในแบบ Cosine
p	คือจุดใด ๆ
q	คือจุดใด ๆ
i	คือจำนวนมิติของข้อมูล

แสดงตัวอย่างการใช้มาตรวัดระยะทาง Cosine Distance

กำหนดให้ $p = \begin{bmatrix}10\\15\end{bmatrix}$ $q = \begin{bmatrix}20\\25\end{bmatrix}$ และคำนวณหาระยะทางโดยใช้ Cosine Distance

$$\begin{aligned} d_3(p,q) &= 1 - \frac{\sum_{i=1}^{n} p_i * q_i}{\sqrt{\sum_{i=1}^{n}(p_i)^2} * \sqrt{\sum_{i=1}^{n}(q_i)^2}} \\ &= 1 - \frac{(10)(20) + (15)(25)}{\sqrt{(10)^2 + (15)^2} * \sqrt{(20)^2 + (25)^2}} \\ &= 1 - \frac{575}{\sqrt{100 + 225} * \sqrt{400 + 625}} \\ &= 1 - \frac{575}{577.17} \\ &= 0.000376 \end{aligned}$$

2.3.4 มาตรวัด Correlation Distance

Correlation Distance (Pearson, 1895) คือมาตรวัดทางสถิติที่ใช้วัดความอิสระต่อกันของค่าสองค่าหรือ เวกเตอร์ 2 เวกเตอร์ใด ๆ โดยค่า Correlation จะมีค่าอยู่ระหว่าง 0 ถึง 1 ซึ่งวัดได้จากค่าความแปรปรวน หรือส่วนเบี่ยงเบนมาตรฐาน ซึ่งสามารถคำนวณได้ตามสมการที่ 2-5 ถึง 2-8

$$d_4(p,q) = 1 - \frac{\mathrm{cov(p,q)}}{\mathrm{std(p)} * \mathrm{std(q)}} \quad (2\text{-}5)$$

$$\mathrm{cov}(\mathrm{p},\mathrm{q}) = \sum_{\mathrm{j}=1}^{\mathrm{k}}(\mathrm{p_j} - \bar{\mathrm{p}}) * (\mathrm{q_j} - \bar{\mathrm{q}}) \quad (2\text{-}6)$$

$$\mathrm{std}(\mathrm{p}) = \sqrt{\frac{1}{\mathrm{k}}\sum_{\mathrm{j}=1}^{\mathrm{k}}(\mathrm{p_j} - \bar{\mathrm{p}})^2} \quad (2\text{-}7)$$

$$\bar{\mathrm{p}} = \frac{1}{\mathrm{k}}\sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{p_j} \quad (2\text{-}8)$$

$d_4(p,q)$	คือระยะทางจากจุด p ไปยังจุด q วัดในแบบ Correlation
p	คือจุดใด ๆ
q	คือจุดใด ๆ
k	คือจำนวนมิติของข้อมูล

ตัวอย่างการใช้มาตรวัดระยะทาง Correlation Distance

กำหนดให้ $p = \begin{bmatrix}10\\15\end{bmatrix}$ $q = \begin{bmatrix}20\\25\end{bmatrix}$ และคำนวณหาระยะทางโดยใช้ Correlation Distance

$$\mathrm{cov}(\mathrm{p},\mathrm{q}) = \sum_{\mathrm{j}=1}^{\mathrm{k}}(\mathrm{p_j} - \bar{\mathrm{p}}) * (\mathrm{q_j} - \bar{\mathrm{q}})$$

$$= \left(10 + \frac{1}{2}(-10 - 15)\right) * \left(20 + \frac{1}{2}(-20 - 25)\right) +$$

$$\left(15 + \frac{1}{2}(-10 - 15)\right) * \left(25 + \frac{1}{2}(-20 - 25)\right)$$

$$= 12.5$$

$$\mathrm{std}(\mathrm{p}) = \sqrt{\frac{1}{\mathrm{k}}\sum_{\mathrm{j}=1}^{\mathrm{k}}(\mathrm{p_j} - \bar{\mathrm{p}})^2}$$

$$= \sqrt{\left(10 + \frac{1}{2}(-10 - 15)\right)^2 + \left(20 + \frac{1}{2}(-20 - 25)\right)^2}$$

$$= \sqrt{\left(10 + (-12.5)\right)^2 + \left(20 + (-22.5)\right)^2}$$

$$= \sqrt{6.25 + 6.25}$$

$$= \sqrt{12.5}$$

$$\text{std(q)} = \sqrt{\left(15 + \frac{1}{2}(-10 - 15)\right)^2 + \left(25 + \frac{1}{2}(-20 - 25)\right)^2}$$

$$= \sqrt{(15 + (-12.5))^2 + (25 + (-22.5))^2}$$

$$= \sqrt{6.25 + 6.25}$$

$$= \sqrt{12.5}$$

$$d_4(p, q) = 1 - \frac{\text{cov(p, q)}}{\text{std(p)} * \text{std(q)}}$$

$$= 1 - \frac{12.5}{\sqrt{12.5} * \sqrt{12.5}}$$

$$= 1 - 1$$

$$= 0$$

ตารางที่ 2.1 เปรียบเทียบมาตรวัดต่าง ๆ โดยสรุปจากตัวอย่างข้างต้น

No.	Type of Distance	Distance Value
1	Euclidean Distance	14.142
2	City Block Distance	20
3	Cosine Distance	0.000376
4	Correlation Distance	0

การเปรียบเทียบมาตรวัดต่าง ๆ โดยสรุปจากการคำนวณดังตารางที่ 2.1 พบว่าค่าของระยะทางในแต่ละมาตรวัดแตกต่างกันออกไป ตามการคำนวณจากสมการที่ต่างกันของแต่ละมาตรวัดระยะทาง ซึ่งในงานวิจัยนี้ได้ทดสอบทั้ง 4 มาตรวัดเพื่อประกอบการดำเนินงานวิจัย

2.4 การทดสอบการกระจายหลายตัวแปรแบบมาร์เดีย

การทดสอบการกระจายหลายตัวแปรแบบมาร์เดีย (Mardia's Multivariate Normality Test) เกิดขึ้นในปี ค.ศ. 1970 มาร์เดียได้เสนอการทดสอบการกระจายแบบปกติหลายตัวแปรจากพื้นฐานการวัดความเบ้และความโด่งของตัวอย่าง $\{x_1, \dots, x_n\}$ สำหรับ k มิติ โดยงานวิจัยนี้ใช้ค่าการวัดความเบ้ของมาร์เดีย (Mardia, 1970) ซึ่งค่าความเบ้ของมาร์เดียสามารถคำนวณได้จากสมการที่ 2-10 ถึง 2-13

$$B_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^3 \quad (2\text{-}10)$$

$$m_{ij} = (x_i - \bar{x})' S^{-1} (x_j - \bar{x}) \quad (2\text{-}11)$$

$$S = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})' \quad (2\text{-}12)$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2\text{-}13)$$

$B_{1,p}$	คือ Mardia's of Multivariate Skew หรือความเบ้ของมาร์เดีย
m_{ij}	คือ ตัวแปรสำหรับลดรูปสมการ
S^{-1}	คือ Covariance Matrix
$\bar{x}$	คือ ค่าเฉลี่ย

สำหรับข้อมูลกลุ่มตัวอย่างขนาดเล็กที่นำมาทดสอบอาจมีความผิดพลาดเกิดขึ้น ดังนั้นมาร์เดียจึงเสนอวิธีการแก้ไข โดยกำหนดให้กลุ่มตัวอย่างที่นำมาทดสอบไม่ควรเกิน 20 กลุ่มตัวอย่างเพื่อควบคุมความผิดพลาด (Mardia, 1974)

2.5 ระบบแนะนำ

ระบบแนะนำ (Recommendation System) คือ ระบบที่สร้างทางเลือกต่าง ๆ ให้ผู้ใช้ โดยที่ผู้ใช้ไม่จำเป็นต้องวิเคราะห์ข้อดี ข้อเสียเอง ทำให้เวลาในส่วนของการค้นหาข้อมูลและวิเคราะห์ข้อมูลนั้นน้อยลง โดยการอาศัยระบบแนะนำเป็นส่วนช่วยสนับสนุนในการตัดสินใจ ระบบแนะนำในปัจจุบันถูกนำไปประยุกต์ในในส่วนต่าง ๆ อย่างแพร่หลาย อาทิ ธุรกรรมทางพาณิชย์ต่าง ๆ เช่น การแนะนำการเลือกซื้อรถยนต์ การแนะนำสถานที่ท่องเที่ยว การแนะนำการเลือกซื้อคอมพิวเตอร์ เป็นต้น นอกจากนี้ระบบแนะนำยังสามารถให้คำแนะนำในประเด็นอื่น ๆ ได้ เช่น การใช้งานอัลกอริทึม การกำหนดพารามิเตอร์ที่เกี่ยวข้อง ทั้งนี้ระบบการแนะนำอาจไม่ใช่วิธีการที่ดีที่สุด แต่

เป็นเพียงการสร้างทางเลือกให้แก่ผู้ใช้และช่วยในการสนับสนุนการตัดสินใจให้แก่ผู้ใช้ (นลินี โสพัศสถิตย, 2555) เทคนิคที่ได้รับความนิยมสำหรับระบบแนะนำมีอยู่ 2 เทคนิคคือ วิธีการใช้เนื้อหา (Content-based) และวิธีการกรองแบบร่วมมือ (Collaborative Filtering) ลักษณะของทั้ง 2 เทคนิคสามารถอธิบายได้ดังนี้

วิธีการใช้เนื้อหา คือการที่เลือกแนะนำสิ่งต่าง ๆ แก่ผู้ที่ต้องการคำแนะนำ โดยพิจารณาจากอดีตที่ผ่านมา ว่าเคยเลือกอะไรมีลักษณะอย่างไร การแนะนำก็จะแนะนำในสิ่งที่คล้ายคลึงกับสิ่งที่เคยถูกเลือกนั้น ตัวอย่างเช่น

นายสมชายชอบร่วมงานที่อยู่ใกล้ทะเล เมื่อมีงานเลี้ยง 2 งานที่จัดขึ้นพร้อมกัน เชิญให้นายสมชายเข้าร่วมงาน ทำให้นายสมชายต้องตัดสินใจในการเลือกว่าจะเข้าร่วมงานเลี้ยงใด การแนะนำก็จะพิจารณาจากในอดีตที่นายสมชายชอบร่วมงานที่อยู่ใกล้ทะเล เมื่อ 1 ใน 2 งานเลี้ยงนั้นมีลักษณะคล้ายคลึงกับงานเลี้ยงที่นายสมชายชอบในอดีต งานเลี้ยงนั้นก็จะถูกแนะนำให้แก่นายสมชาย

วิธีการกรองแบบร่วมมือ คือการที่เลือกแนะนำสิ่งต่าง ๆ แก่ผู้ที่ต้องการคำแนะนำโดยพิจารณาจากบุคคลอื่นที่มีความชอบคล้ายคลึงกับผู้ที่ต้องการนำแนะนำ ตัวอย่างเช่น

นายสมชายชอบเที่ยวทะเลและตกปลา นายดีชอบเที่ยวทะเล เมื่อนายดีต้องเลือกทำกิจกรรมอย่างใดอย่างหนึ่งระหว่างตกปลา และพายเรือ การแนะนำก็จะพิจารณาจากบุคคลอื่นที่มีความชอบคล้ายคลึงกัน โดยในตัวอย่างนี้คือนายสมการ การแนะนำจึงแนะนำกิจกรรมตกปลาแก่นายดี

ในงานวิจัยนี้ได้นำเสนอเกี่ยวกับการกำหนดค่า เค ของเคเนียร์เรสเนเบอร์ และการกำหนดมาตรวัดระยะทาง ซึ่งมีลักษณะในการนำเสนอเชิงแนะนำเพื่อเป็นทางเลือกในการตัดสินใจแก่ผู้ที่สนใจและต้องการใช้งาน เพื่อช่วยลดระยะเวลาในการวิเคราะห์ข้อดีและข้อเสียในการเลือกค่า เค ของเคเนียร์เรสเนเบอร์และการกำหนดมาตรวัดระยะทาง ทั้งนี้การแนะนำอาจไม่ใช่คำตอบที่ดีที่สุด แต่เป็นการสร้างทางเลือกให้แก่ผู้ที่สนใจและต้องการใช้งาน

2.6 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการหาค่า เค ของเคเนียร์เรสเนเบอร์ โดยบางส่วนของงานวิจัยที่เกี่ยวข้องมักใช้เทคนิคหรืออัลกอริทึมที่คิดค้นขึ้น นอกจากนี้บางงานวิจัยยังใช้เทคนิคด้านสถิติเพื่อช่วยหาค่า เค อีกด้วย โดยงานวิจัยที่เกี่ยวข้องต่อไปนี้ใช้ข้อมูลทางการแพทย์เพื่อทดสอบผลทั้งสิ้น

D. J. Hand และ V. Vinciotti (2003) ได้ศึกษาเกี่ยวกับการเลือกค่า เค สำหรับข้อมูลที่มีคลาสจำนวน 2 คลาส โดยลักษณะเป็น Unbalanced Classes คือมีจำนวนข้อมูลในคลาสใดคลาสหนึ่งมากกว่าอีกคลาสหนึ่งมาก ๆ โดยในงานวิจัยนี้ได้กล่าวถึงการใช้ Bayes Theorem ในการหาค่า เค ที่เหมาะสมสำหรับข้อมูลสองคลาสที่มีลักษณะ Unbalanced Classes และมีการเปรียบเทียบผลลัพธ์ใน

การใช้กราฟในการแสดงข้อมูลเกี่ยวกับค่าความแม่นกับค่า เค ซึ่งในงานวิจัยนี้แสดงเพียงค่า เค ที่เริ่มจาก 1 ถึง 100 เท่านั้น

A. Ghosh (2006) ได้ศึกษาเกี่ยวกับการเลือกค่า เค ที่ดีที่สุดของเคเนียร์เรสเนเบอร์โดยในงานวิจัยนี้มีการใช้เทคนิคด้านสถิติเกี่ยวกับความน่าจะเป็นโดย Bayes Theorem ซึ่งได้เสนอเป็นทฤษฎี และมีการทดสอบกับเทคนิคอื่นที่ใช้หาค่า เค ด้วยกันอีกสองเทคนิคคือ Likelihood Cross-Validation (LCV) และ Leave-One-Out Cross-Validation (CV-Class) โดยทั้ง 3 เทคนิคได้ใช้ทดสอบกับข้อมูลต่าง ๆ ทั้งที่เป็นเป็น 2 คลาส และมากกว่า 2 คลาส ซึ่งผลการทดสอบเป็นที่น่าพอใจในข้อมูลที่มีลักษณะ 2 คลาส เทคนิคของเขาให้ค่าความผิดพลาดต่ำที่สุดหากเปรียบเทียบกับเทคนิค LCV และ CV-Class แต่เมื่อทดสอบกับข้อมูลที่มีลักษณะ 3 คลาส เทคนิคนี้มีประสิทธิภาพในระดับปานกลาง

C. Hulett และคณะ (2012) ได้ศึกษาเกี่ยวกับการเลือกค่า เค ของเคเนียร์เรสเนเบอร์ใน Instance-Based Learning ซึ่งมีลักษณะไม่ต่างจาก Lazy-Learing คือเปรียบเทียบข้อมูลใหม่กับข้อมูลเดิม หากคล้ายคลึงกันก็ถูกจัดให้เป็นไปตามลักษณะข้อมูลเดิม โดยทีมวิจัยนี้ได้เสนอเทคนิคใหม่ที่มีชื่อว่า IBcs Algorithm โดยพิจารณาเลือก Instance ที่เหมาะสมมาเป็นค่า เค ที่เหมาะสมเมื่อเปรียบเทียบ IBcs กับเคเนียร์เรสเนเบอร์ในด้านความแม่นซึ่งโดยรวมแล้วค่า เค ที่ดีที่สุดที่ให้ค่าความแม่นสูงจากเคเนียร์เรสเนเบอร์ก็ยังสูงกว่าค่าที่ได้จาก IBcs แต่ก็ห่างกันไม่มาก และเมื่อเขาเปรียบเทียบค่า เค บางค่ากับ IBcs พบว่า IBcs ให้ค่าความแม่นสูงใกล้เคียงกับค่า เค บางค่าที่ความแม่นสูงที่สุด นอกจากนี้ยังเปรียบเทียบเวลากับ Cross-Validation อีกด้วย ซึ่ง IBcs ใช้เวลาต่ำสุด

S. A. Medjahed และคณะ (2013) ได้ศึกษาเกี่ยวกับการวินิจฉัยโรคมะเร็งซึ่งเป็นปัญหามากในการศึกษาทางการแพทย์เกี่ยวกับโรคนี้ ซึ่งนักวิจัยหลายคนร่วมกันปรับปรุงประสิทธิภาพมาอย่างยาวนาน โดยในงานวิจัยนี้ได้ใช้เคเนียร์เรสเนเบอร์ในการวินิจฉัยทางการแพทย์โดยความน่าสนใจของเคเนียร์เรสเนเบอร์ที่พบคือระยะทางและค่า เค ส่งผลที่แตกต่างกัน ในการกำหนดค่า เค โดยทีมวิจัยนี้ได้ใช้ เทคนิคใหม่หลายเทคนิคเพื่อการพิจารณาเลือกค่า เค ที่เหมาะสม มาใช้กับข้อมูลโรคมะเร็งเต้านมและนอกจากนี้มาตรวัดระยะทางที่เลือกใช้ มีหลากหลายประเภทเช่นกัน

G. Bhattacharya และคณะ (2014) ได้ศึกษาเกี่ยวการการทดสอบโดยเทคนิคพิเศษในการประมาณค่า เค ของเคเนียร์เรสเนเบอร์ซึ่งในการศึกษาครั้งนี้ได้กำหนดค่า เค บางส่วนเพื่อใช้เปรียบเทียบกับการทดสอบ คือ เค ที่เป็น 1, 3, 5, 7 และ $\sqrt{n}$ โดย n คือจำนวนข้อมูลทั้งหมด ในที่นี้การทดสอบใช้กับมาตรวัด Euclidean และ City Block เทคนิคในการทดสอบเกี่ยวกับการให้น้ำหนักของ Huness (Tomasev et al., 2012) และทำการเปรียบเทียบกับค่า เค ที่ถูกกำหนดก่อนหน้านี้คือ 1, 3, 5, 7 และ $\sqrt{n}$ ซึ่งเทคนิคให้ค่าความแม่นโดยรวมค่อนข้างสูงไม่ว่าจะเปรียบเทียบโดยมาตรวัด

Euclidean หรือ City Block ก็ตาม นอกจากนี้ยังเปรียบเทียบ กับอัลกอริทึมอื่น ๆ เช่น hw-kNN, h-kNN และ dw h-FNN เป็นต้น

J. W. Yoon และ N. Friel (2015) ได้ศึกษาเกี่ยวกับ Probabilistic k-Nearest Neighbor หรือ PKNN ซึ่งเป็นเคเนียร์เรสเนเบอร์ในลักษณะที่มีความน่าจะเป็นเข้ามาเกี่ยวข้องและได้มีการพัฒนาโดยนำ ศาสตร์ทางด้านสถิติเข้ามาคำนวณร่วมด้วย ไม่ว่าจะเป็น Pseudo-Likelihood, Markov Chain Monte Carlo, Bayesian Model Selection และ Integrated Nested Laplace Approximation ซึ่งหลังจากการคำนวณในครั้งแรกเขาได้พัฒนาการคำนวณอีกลักษณะหนึ่งขึ้นโดยมีการปรับเปลี่ยนตัวแปรจากการคำนวณครั้งแรกไม่มากนัก แต่ก็ทำให้ผลที่ได้ออกมาดีขึ้น

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่าการให้ความสำคัญกับการกำหนดค่า เค ของเคเนียร์เรสเนเบอร์นั้นมีผลกับค่าความแม่นในการจำแนกประเภทของเคเนียร์เรสเนเบอร์มาก นอกจากนี้มาตรวัดระยะทางที่ใช้ในการจำแนกประเภทก็มีผลให้ค่าความแม่นแตกต่างกันไปเช่นกัน สาระสำคัญในงานวิจัยนี้เมื่อเปรียบเทียบกับงานวิจัยอื่นสรุปได้ดังตารางที่ 2.2

ตารางที่ 2.2 เปรียบเทียบโดยสรุปงานวิจัยที่เกี่ยวข้องกับการเลือกค่า เค ที่เหมาะสมของ k-Nearest Neighbor กับข้อมูลทางการแพทย์

กระบวนการทำงาน	**งานวิจัยที่เกี่ยวข้อง**						
	ก	**ข**	**ค**	**ง**	**จ**	**ฉ**	**ช***
อัลกอริทึมที่เกี่ยวข้อง							
Nearest Neighbor Classification	✓	✓	✓	✓	✓	✓	✓
Bayes Theorem	✓	✓					
Cross-Validation		✓	✓		✓	✓	
อัลกอริทึมอื่น ๆ			✓	✓	✓	✓	✓
เกณฑ์การประเมินประสิทธิภาพ							
Accuracy	✓	✓	✓	✓	✓	✓	✓
t-Test					✓		
Bayesian Method	✓	✓					
มาตรวัดระยะทาง							
Euclidean Distance	✓	✓	✓	✓	✓		✓
City Block (or Manhattan) Distance			✓	✓	✓		✓
Cosine Distance				✓			✓

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 2.2 เปรียบเทียบโดยสรุปงานวิจัยที่เกี่ยวข้องกับการเลือกค่า เค ที่เหมาะสมของ k-Nearest Neighbor กับข้อมูลทางการแพทย์ (ต่อ)

กระบวนการทำงาน	งานวิจัยที่เกี่ยวข้อง						
	ก	ข	ค	ง	จ	ฉ	ช*
มาตรวัดระยะทาง							
Correlation Distance				✓			✓
มาตรวัดระยะทางอื่น ๆ		✓	✓			✓	
ข้อมูลที่ใช้ในงานวิจัย							
Medical Datasets	✓	✓	✓	✓	✓	✓	✓
Other Datasets		✓	✓		✓	✓	
ขอบเขตของงานวิจัย							
วิจัยเพื่อทดสอบประสิทธิภาพ	✓	✓	✓	✓	✓	✓	✓
วิจัยเพื่อเสนอแนวคิดใหม่	✓	✓	✓	✓	✓	✓	✓
มีการประยุกต์ใช้กับข้อมูลจริง	✓	✓	✓	✓	✓	✓	✓

หมายเหตุ งานวิจัยที่เกี่ยวข้องประกอบด้วย

ก แทนงานวิจัยของ D. J. Hand และ V. Vinciotti (2003)

ข แทนงานวิจัยของ A. Ghosh (2006)

ค แทนงานวิจัยของ C. Hulett และคณะ (2012)

ง แทนงานวิจัยของ S. A. Medjahed และคณะ (2013)

จ แทนงานวิจัยของ G. Bhattacharya และคณะ (2014)

ฉ แทนงานวิจัยของ J. W. Yoon และ N. Friel (2015)

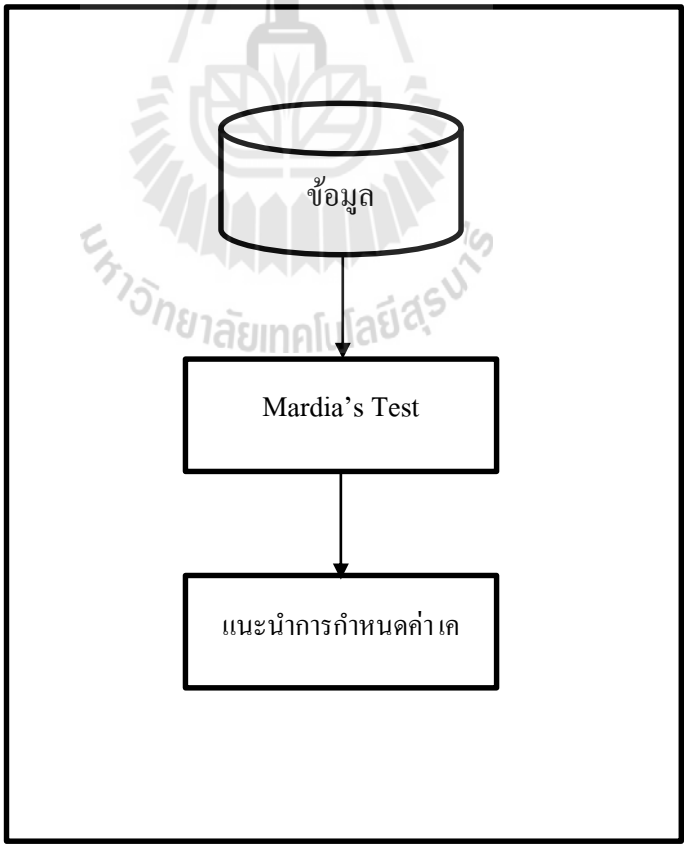
ช* แทนวิทยานิพนธ์ฉบับนี้

บทที่ 3
วิธีดำเนินงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อแนะนำค่า เค ที่เหมาะสมของอัลกอริทึมเคเนียร์เรสเนเบอร์ ซึ่งในบทนี้จะกล่าวถึงข้อมูลที่ใช้ในงานวิจัย วิธีดำเนินงานวิจัย เครื่องมือที่ใช้ในงานวิจัย และกระบวนการต่าง ๆ ของงานวิจัย โดยมีรายละเอียดดังนี้

3.1 กรอบแนวคิดของงานวิจัย

แนวคิดหลักของงานวิจัยนี้ คือการแนะนำเกี่ยวกับการกำหนดค่า เค ของเคเนียร์เรสเนเบอร์และมาตรวัดระยะทาง โดยการพิจารณาจากลักษณะการกระจายของข้อมูล ซึ่งวัดโดยการทดสอบแบบมาร์เดีย



รูปที่ 3.1 การแนะนำในการกำหนดค่า เค

จากรูปที่ 3.1 เป็นกรอบแนวคิดในการแนะนำค่า เค สำหรับผู้ใช้งาน โดยสามารถแบ่งออกเป็น 3 ขั้นตอนหลัก ๆ คือ

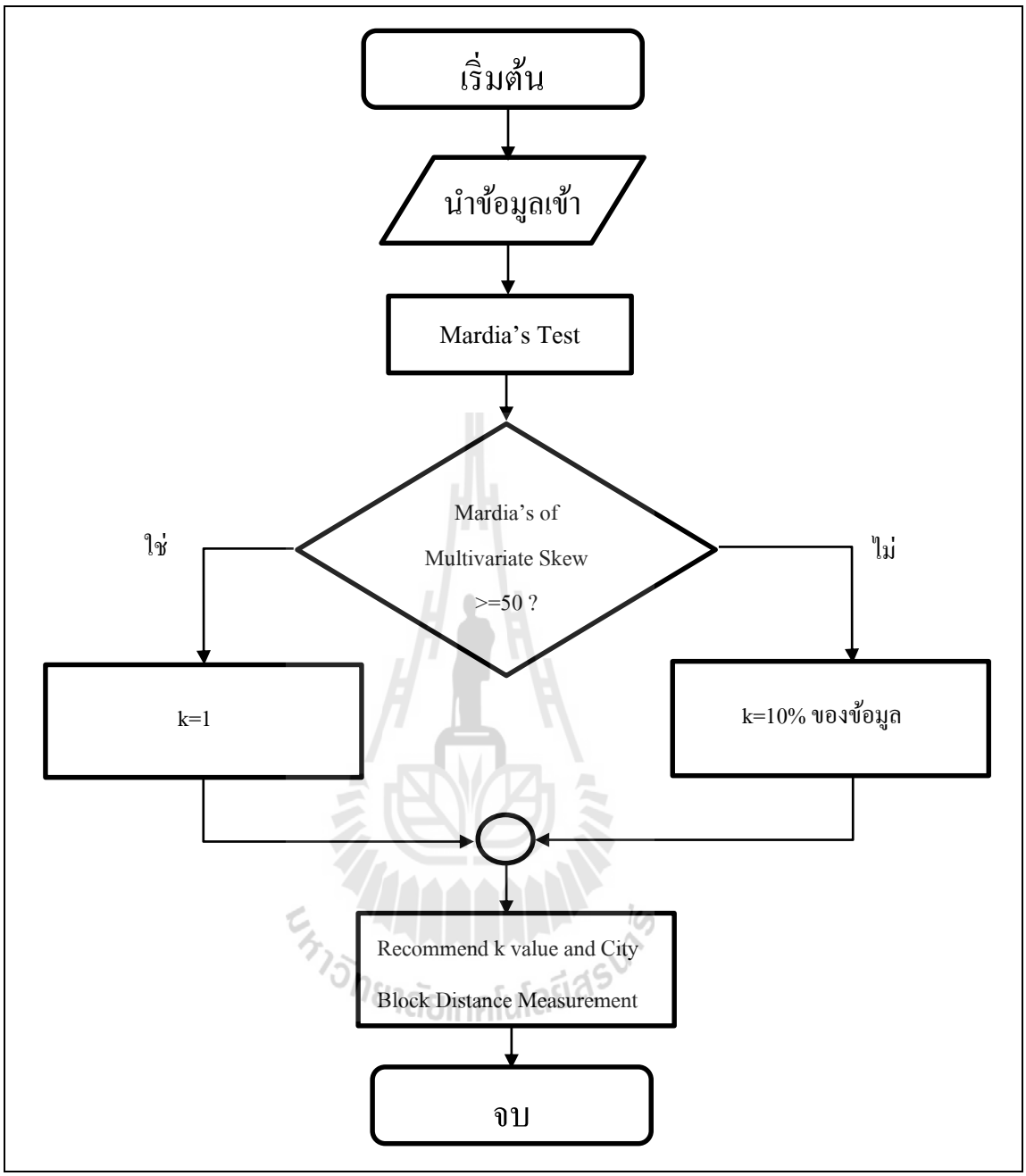
1. เลือกข้อมูลที่ต้องการ
2. นำข้อมูลที่ต้องการวัดการกระจายโดยการทดสอบแบบมาร์เดีย
3. พิจารณาจากค่าของมาร์เดียเพื่อใช้ในการกำหนดค่า เค และมาตรวัดระยะทาง

3.2 การออกแบบอัลกอริทึม

3.2.1 อัลกอริทึมการเลือกค่า เค ที่เหมาะสมกับข้อมูลทางการแพทย์

การออกแบบอัลกอริทึมเพื่อใช้ในการเลือกค่า เค ที่เหมาะสมในงานวิจัยนี้ได้เสนอการใช้ค่าความเบ้ของมาร์เดีย เพื่อช่วยในการตัดสินใจโดยวัดค่าความเบ้ของมาร์เดียจากชุดข้อมูลทดสอบ โดยที่ทำการสุ่มตัวอย่างข้อมูล 20 ข้อมูล 3 ครั้งจากข้อมูลทดสอบ และหาค่าเฉลี่ยค่าความเบ้ของมาร์เดีย เพื่อใช้เลือกค่า เค ที่เหมาะสม โดยมีขั้นตอนดังรูปที่ 3.3

มหาวิทยาลัยเทคโนโลยีสุรนารี



รูปที่ 3.2 ผังงานแสดงขั้นตอนการเลือกค่า เค ที่เหมาะสม

จากรูปที่ 3.2 เป็นขั้นตอนในการเลือกค่า เค ที่เหมาะสม ซึ่งขั้นตอนแรกนำข้อมูลมาผ่านขั้นตอน Mardia's of Multivariate Skew >= 50 ? โดยในขั้นตอนนี้เป็นการวัดค่าความเบ้ของมาร์เดียกับชุดข้อมูล โดยจะทำการสุ่มตัวอย่างออกมาเพื่อวัดค่า การสุ่มตัวอย่างนั้นจะสุ่ม 3 ครั้ง ในแต่ละครั้งของการสุ่มจะสุ่มข้อมูลมาเพียง 20 ข้อมูล และทำการวัดค่าความเบ้ของมาร์เดียใน 3 ครั้งนั้น และ

คำนวณหาค่าเฉลี่ยค่าความเบ้ของมาร์เดียเพื่อใช้เป็นเกณฑ์ในการเลือกค่า เค ที่เหมาะสม ซึ่งหากข้อมูลมีค่าเฉลี่ยความเบ้ของมาร์เดียมากกว่าหรือเท่ากับ 50 ในขั้นตอนการเลือกค่า เค และมาตรวัดระยะทางสามารถเลือกค่า เค เป็น 1 กับมาตรวัดระยะทาง City Block ได้ทันที และหากข้อมูลมีค่าเฉลี่ยความเบ้ของมาร์เดียมีค่าน้อยกว่า 50 สามารถเลือกค่า เค เป็น 10% ของจำนวนข้อมูลทั้งหมดกับมาตรวัดระยะทาง City Block หลังจากนั้นสามารถใช้การแนะนำที่ได้จากอัลกอริทึมจากงานวิจัยนี้ในการจำแนกข้อมูลด้วยเคเนียร์เรสเนเบอร์ เพื่อผลลัพธ์ที่ต้องการต่อไป

3.2.2 การแนะนำ

การแนะนำเกี่ยวกับการกำหนดค่า เค และมาตรวัดระยะทางของเคเนียร์เรสเนเบอร์เพื่อให้ได้ค่าความแม่นที่สูง คือหัวใจหลักของงานวิจัยนี้ ซึ่งการแนะนำจะใช้อัลกอริทึมจากงานวิจัยนี้เพื่อแนะนำค่า เค และมาตรวัดระยะทางที่ใช้กับเคเนียร์เรสเนเบอร์ โดยมีลักษณะขั้นตอนในกระบวนการแนะนำดังต่อไปนี้

1. ผู้ใช้เตรียมข้อมูล
2. สุ่มตัวอย่างจากข้อมูลมา 20 ตัวอย่าง จำนวน 3 ครั้ง
3. ใช้ข้อมูลที่สุ่มมาทั้ง 3 ครั้งเพื่อหาค่าเฉลี่ยของการทำการทดสอบแบบมาร์เดีย
4. บันทึกค่าเฉลี่ย Mardia’s of Multivariate Skew
5. พิจารณาค่าเฉลี่ย Mardia’s of Multivariate Skew ว่ามากกว่าหรือเท่ากับ 50 หรือไม่
6. ค่าเฉลี่ย Mardia’s of Multivariate Skew จะเป็นตัวแนะนำผู้ใช้ว่า ควรจะเลือกใช้ค่า เค และมาตรวัดระยะทาง โดยถ้าค่าเฉลี่ย Mardia’s of Multivariate Skew มากกว่าหรือเท่ากับ 50 เลือกค่า เค เป็น 1 และใช้มาตรวัดระยะทาง City Block หากข้อมูลมีค่าเฉลี่ย Mardia’s of Multivariate Skew น้อยกว่า 50 สามารถเลือกค่า เค เป็น 10% ของจำนวนข้อมูลทั้งหมดกับมาตรวัดระยะทาง City Block
7. เมื่อผู้ใช้ได้คำแนะนำหลังจากนั้นสามารถกำหนดค่า เค เพื่อใช้งานกับเคเนียร์เรสเนเบอร์ได้ทันที

ตัวอย่างการแนะนำโดยอัลกอริทึมจากงานวิจัยนี้เมื่อผู้ใช้มีข้อมูลที่ต้องการ ดังตารางที่ 3.1 ซึ่งเป็นการสมมุติเหตุการณ์ เพื่อจำลองการใช้งานอัลกอริทึมจากงานวิจัยนี้

ตารางที่ 3.1 ตัวอย่างข้อมูล

Column 1	Column 2
2	3
1	5
2	7
4	1

ในขั้นตอนถัดไปหากดำเนินการตามอัลกอริทึมจากงานวิจัยนี้จำเป็นต้องสุ่มข้อมูลมา 20 ข้อมูล ซึ่งในตัวอย่างที่จะแสดงต่อไปนี้ เป็นเพียงการสุ่ม 3 ข้อมูลเท่านั้น ทั้งนี้เพื่อการคำนวณที่ง่ายและดูไม่ซับซ้อนจนเกินไป ข้อมูลสุ่มครั้งที่ 1 เป็นดังตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างข้อมูลสุ่ม

Column 1	Column 2
2	3
1	5
2	7

หลังจากสุ่มข้อมูลตัวอย่าง สามารถนำข้อมูลข้างต้นมาทำการคำนวณเพื่อหาค่าเฉลี่ย Mardia's of Multivariate Skew ได้ดังต่อไปนี้

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{n} x_i}{n} \\ &= \frac{2+1+2+3+5+7}{6} \\ &= \frac{20}{6} \\ &= 3.33\end{aligned}$$

$$
\begin{aligned}
S = & \frac{1}{n}\sum_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})' \\
= & \frac{1}{3}[(3-3.33)(3-3.33)+ \\
& (5-3.33)(5-3.33)+ \\
& (7-3.33)(7-3.33)] \\
= & 22.14
\end{aligned}
$$

$$
\begin{aligned}
B_{1,p} = & \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}[(x_i-\bar{x})'S^{-1}(x_j-\bar{x})]^3 \\
= & \tfrac{1}{3^2}[(2-3.33)(22.14)^{-1}(3-3.33)]^3+ \\
& [(2-3.33)(22.14)^{-1}(5-3.33)]^3+ \\
& [(2-3.33)(22.14)^{-1}(7-3.33)]^3+ \\
& [(1-3.33)(22.14)^{-1}(3-3.33)]^3+ \\
& [(1-3.33)(22.14)^{-1}(5-3.33)]^3+ \\
& [(1-3.33)(22.14)^{-1}(7-3.33)]^3+ \\
& [(2-3.33)(22.14)^{-1}(3-3.33)]^3+ \\
& [(2-3.33)(22.14)^{-1}(5-3.33)]^3+ \\
& [(2-3.33)(22.14)^{-1}(7-3.33)]^3 \\
= & -0.647
\end{aligned}
$$

จากการคำนวณข้างต้นเป็นเพียงการสุ่มครั้งที่ 1 สำหรับการสุ่มครั้ง 2 และครั้งที่ 3 การคำนวณหาค่า Mardia's of Multivariate Skew มีลักษณะเหมือนกัน และเมื่อสามารถหาค่า Mardia's of Multivariate Skew จากการสุ่มข้อมูลทั้ง 3 ครั้งได้แล้ว นำค่า Mardia's of Multivariate

Skew จากทั้ง 3 การสุ่มมาเฉลี่ย เนื่องจากตัวอย่างการคำนวณข้างต้นแสดงการหาค่า Mardia's of Multivariate Skew เพียงการสุ่มครั้งแรก หากการสุ่มครั้งที่ 2 และครั้งที่ 3 มีค่า Mardia's of Multivariate Skew เป็น 4.234 และ 3.143 ตามลำดับ การเฉลี่ยค่า Mardia's of Multivariate Skew สามารถทำได้ดังต่อไปนี้

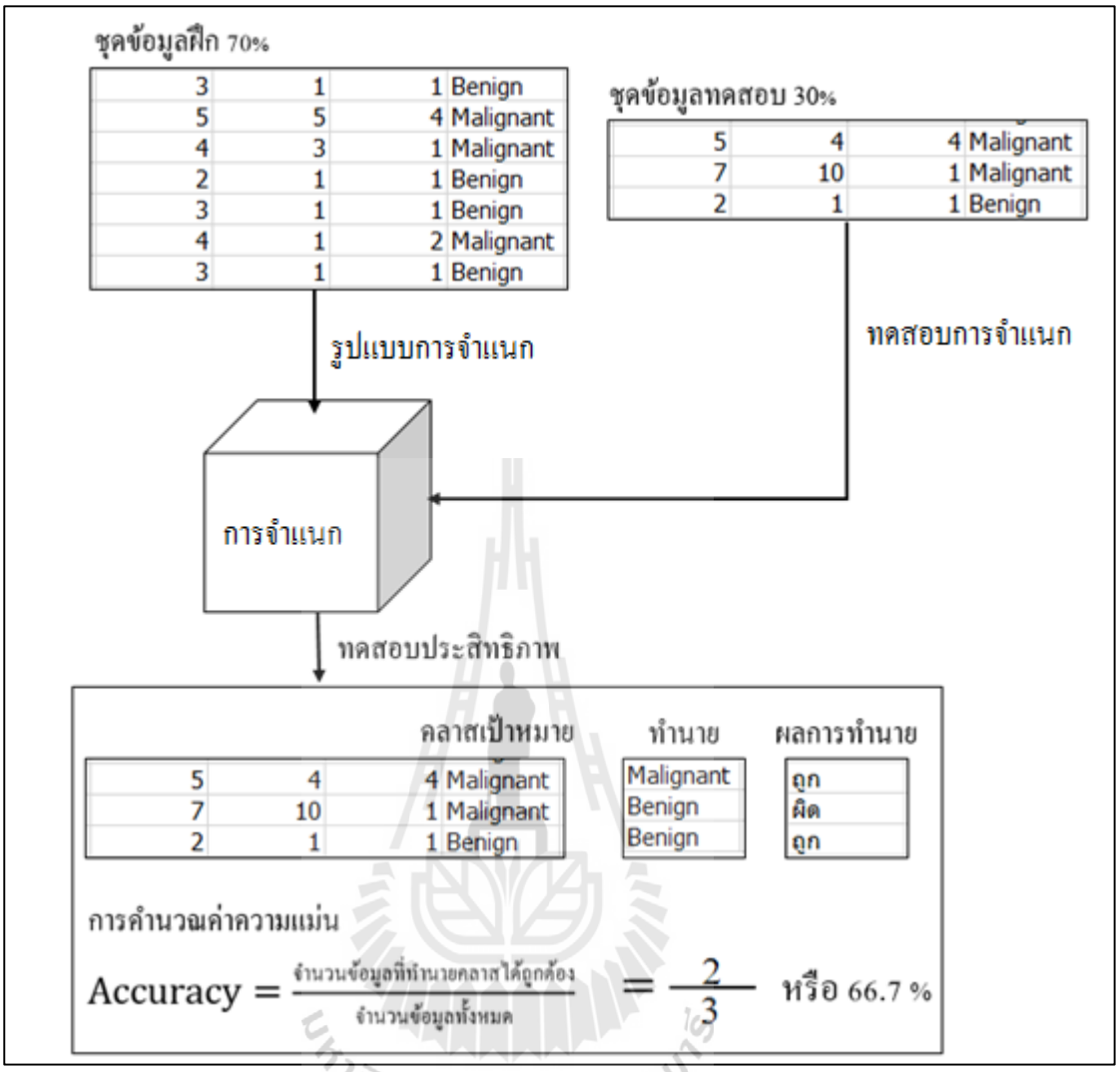
$$\text{ค่าเฉลี่ย Mardia's of Multivariate Skew} = \frac{-0.647 + 4.234 + 3.143}{3}$$
$$= 2.24$$

เมื่อได้ค่าเฉลี่ย Mardia's of Multivariate Skew พิจารณาว่ามากกว่าหรือเท่ากับ 50 หรือไม่ ปรากฏว่าผลการคำนวณที่ได้มีค่าน้อยกว่า การกำหนดค่า เค สำหรับข้อมูลตัวอย่างนี้ จึงถูกแนะนำโดยใช้ เค เท่ากับ 1 และมาตรวัดระยะทาง City Block

3.2.3 การทดสอบประสิทธิภาพของการเลือกค่า เค ที่เหมาะสมกับข้อมูลทางการแพทย์

การทดสอบประสิทธิภาพในการเลือกค่า เค สามารถทำได้โดยการพิจารณาค่าความแม่นจากการจำแนกโดยจะทำการแบ่งข้อมูลออกเป็น 2 ส่วน คือชุดข้อมูลฝึก และชุดข้อมูลทดสอบ โดยจะใช้ชุดข้อมูลฝึกในการสร้างรูปแบบการจำแนกและใช้ข้อมูลทดสอบมาทดสอบผลกับรูปแบบการจำแนกนั้นเพื่อวัดประสิทธิภาพในการทำนาย ซึ่งสามารถทดสอบประสิทธิภาพได้ตามตัวอย่างดังรูปที่ 3.3

มหาวิทยาลัยเทคโนโลยีสุรนารี



รูปที่ 3.3 ตัวอย่างการทดสอบประสิทธิภาพการจำแนก

3.3 เครื่องมือที่ใช้ในงานวิจัย

เครื่องมือที่ใช้ในงานวิจัยนี้ ใช้คอมพิวเตอร์ซึ่งประกอบด้วยฮาร์ดแวร์และซอฟต์แวร์ ดังนี้

ฮาร์ดแวร์

- หน่วยประมวลผล : Intel® Core i3
- หน่วยความจำสำรอง : 500 GB
- หน่วยความจำหลัก : 2 GB
- อุปกรณ์เสริมอื่น ๆ เช่น เมาส์ แป้นพิมพ์ เป็นต้น

ซอฟต์แวร์

- ระบบปฏิบัติการ : Windows 8 Pro
- โปรแกรม MATLAB
- โปรแกรม RStudio

บทที่ 4
การทดสอบและอภิปรายผล

ในบทนี้จะกล่าวถึงการทดลองผลจากอัลกอริทึม และข้อมูลต่าง ๆ ที่ใช้ในการทดสอบ ซึ่งจะมีการแสดงผลการทดลองระหว่างการใช้อัลกอริทึมและไม่ใช้อัลกอริทึมจากงานวิจัยนี้ โดยลักษณะการนำเสนอผลการทดลองและกระบวนการดำเนินงานจะมีหลากหลายลักษณะ เช่น กราฟ ตาราง รูปภาพ เป็นต้น

4.1 ข้อมูลที่ใช้ในการทดสอบ

ในการทดสอบแนวทางการเลือกค่า เค ที่เสนอในงานวิจัยนี้ใช้ข้อมูลทางการแพทย์ ซึ่งข้อมูลที่ใช้เหล่านี้เป็นข้อมูลที่มี 2 คลาสเป้าหมาย หรือ 3 คลาสเป้าหมายในการทดสอบเท่านั้น เนื่องจากข้อมูลทางการวินิจฉัยโรคมักจะมีจำนวนคลาสไม่มากโดยประกอบด้วยข้อมูล โรคมะเร็งเต้านม โรคหัวใจ โรคเบาหวาน และโรคไทรอยด์ จาก UCI Machine Learning Repository และข้อมูลโรคหอบหืด จากโรงพยาบาลมหาราช จังหวัดนครราชสีมา โดยทำการรวบรวมข้อมูล ณ วันที่ 8 พฤศจิกายน พ.ศ. 2557 ซึ่งข้อมูลข้างต้นมีลักษณะดังต่อไปนี้

ตารางที่ 4.1 ข้อมูลโรคมะเร็งเต้านม

ลำดับ	คอลัมน์	ค่า
1	Sample Code Number	หมายเลขผู้ป่วย
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10

ตารางที่ 4.1 ข้อมูลโรคมะเร็งเต้านม (ต่อ)

ลำดับ	คอลัมน์	ค่า
10	Mitoses	1-10
11	Class	Benign, Malignant

จากตารางที่ 4.1 เป็นข้อมูลโรคมะเร็งเต้านม ซึ่งประกอบด้วย 11 คอลัมน์ จำนวนข้อมูลทั้งหมด 699 ข้อมูล ซึ่งในคอลัมน์ Sample Code Number จะไม่ถูกทำมาวิเคราะห์เนื่องจากเป็นเพียงหมายเลขผู้ป่วยซึ่งไม่ส่งผลกับการจำแนก

ตารางที่ 4.2 ข้อมูลโรคหัวใจ

ลำดับ	คอลัมน์	ค่า
1	Age	29-77
2	Sex	Male, Female
3	Chest Pain Type	0, 1, 2, 3
4	Resting Blood Pressure	94-200
5	Serum Cholestoral	126-564
6	Fasting Blood Sugar > 120 mg/dl	0, 1
7	Resting Electrocardiographic Results	0, 1
8	Maximum Heart Rate Achieved	71-202
9	Exercise Induced Angina	0, 1
10	Oldpeak	0-6.2
11	The Slope of The Peak Exercise ST Segment	0, 1, 2
12	Number of Major Vessels	0, 1, 2, 3
13	Thal	0, 1, 2
14	Class	Absence, Presence

จากตารางที่ 4.2 เป็นข้อมูลโรคหัวใจ ซึ่งประกอบด้วย 14 คอลัมน์ จำนวนข้อมูลทั้งหมด 270 ข้อมูล

ตารางที่ 4.3 ข้อมูลโรคเบาหวาน

ลำดับ	คอลัมน์	ค่า
1	Number of Times Pregnant	0-17
2	Plasma Glucose Concentration a 2 Hours in an Oral Glucose Tolerance Test	0-199
3	Diastolic Blood Pressure	0-122
4	Triceps Skin Fold Thickness	0-99
5	2-Hour Serum Insulin	0-846
6	Body Mass Index	0-67.1
7	Diabetes Pedigree Function	0.078-2.42
8	Age	21-81
9	Class	Yes, No

จากตารางที่ 4.3 เป็นข้อมูลโรคเบาหวาน ซึ่งประกอบด้วย 9 คอลัมน์ จำนวนข้อมูลทั้งหมด 768 ข้อมูล

ตารางที่ 4.4 ข้อมูลโรคหอบหืด

ลำดับ	คอลัมน์	ค่า
1	Case Number	หมายเลขผู้ป่วย
2	Age Respondents (Years)	35-64
3	Gender of Respondents	0, 1
4	Highest Education Level	1, 2, 3, 4
5	Marital Status	1, 2, 3, 4, 5
6	Religion	1, 2, 3
7	Smoking	0, 1, 2
8	Exercise	0, 1
9	Weight (kg.)	37.2-113.3
10	Weight (cm.)	141-192
11	Waist (cm.)	57-119
12	Percent Body Fat	11.7-47.6

ตารางที่ 4.4 ข้อมูลโรคหอบหืด (ต่อ)

ลำดับ	คอลัมน์	ค่า
13	PA level	Low, Moderate, High

จากตารางที่ 4.4 เป็นข้อมูลโรคหอบหืด ซึ่งประกอบด้วย 13 คอลัมน์ จำนวนข้อมูลทั้งหมด 677 ข้อมูล

ข้อมูลโรคไทรอยด์ จำนวนข้อมูลทั้งหมด 7200 ข้อมูล โดยมี 22 คอลัมน์ ข้อมูลนี้ได้จากแหล่งข้อมูลมาตรฐาน UCI Machine Learning Repository ซึ่งถูกบริจารในปี ค.ศ. 1987 โดย Peter Turney

จากรูปที่ 4.1 เป็นรูปแสดงตัวอย่างของข้อมูลโรคหอบหืด ซึ่งจะพบว่ามีเพียง 12 คอลัมน์เท่านั้นที่ถูกใช้งาน โดยคอลัมน์ที่ไม่ส่งผลต่อโรคจะถูกตัดทิ้ง เช่น คอลัมน์หมายเลขผู้ป่วย หมายเลขประจำตัวประชาชน หมายเลขโทรศัพท์ เป็นต้น

age	sex	edu	ms	rel	smoking	exercise	wt	ht	wc	bodyfat	PA_level
54	1	2	2	1	0	1	62.4	163	75.5	33	moderate
52	1	3	2	1	1	0	47.5	153	65.5	27.6	low
51	1	3	2	1	0	1	57.1	160	83	31.6	moderate
55	1	3	1	1	0	0	57.7	150	87	38.2	low
57	1	2	2	1	0	0	53.4	158	76	32.2	low

รูปที่ 4.1 ตัวอย่างข้อมูลโรคหอบหืด

ข้อมูลที่ใช้ในการทดสอบจะมีลักษณะเป็นตัวเลข และมีคลาสเป้าหมายเป็นตัวอักษร โดยมีลักษณะข้อมูลตัวอย่าง ดังรูปที่ 4.1 ถึง 4.5 ซึ่งเป็นบางส่วนของข้อมูล ทั้งนี้ในบางข้อมูลจำเป็นต้องมีการจัดการข้อมูลก่อน การดำเนินการทดสอบเช่นข้อมูลที่มีลักษณะที่วัดไม่ได้ เช่น เพศชาย เพศหญิง จึงจำเป็นต้องแปลงข้อมูลนั้น ๆ ให้อยู่ในรูปของ Dummy Variable เพื่อช่วยให้อัลกอริทึมสามารถใช้งานกับข้อมูลได้ โดยที่ลักษณะการแปลงข้อมูลให้อยู่ในรูป Dummy Variable มีลักษณะดังรูปที่ 4.6

130	322	0	1	1	0	109	0	1	2.4	0	1	0	3	0	1	0 sick
115	564	0	1	1	0	160	0	1	1.6	0	1	0	0	0	0	1 healthy
124	261	0	1	0	1	141	0	1	0.3	0	0	1	0	0	0	1 sick
128	263	0	1	0	1	105	1	0	0.2	0	1	0	1	0	0	1 healthy
120	269	0	1	1	0	121	1	0	0.2	0	0	1	1	0	1	0 healthy
120	177	0	1	0	1	140	0	1	0.4	0	0	1	0	0	0	1 healthy

รูปที่ 4.2 ตัวอย่างข้อมูลโรคหัวใจ

6	8	8	1	3	4	3	7	1	Benign
4	1	1	3	2	1	3	1	1	Benign
8	10	10	8	7	10	9	7	1	Malignant
1	1	1	1	2	10	3	1	1	Benign
2	1	2	1	2	1	3	1	1	Benign

รูปที่ 4.3 ตัวอย่างข้อมูลโรคมะเร็งเต้านม

8	183	64	0	0	23.3	0.672	32	Yes
1	89	66	23	94	28.1	0.167	21	No
0	137	40	35	168	43.1	2.288	33	Yes
5	116	74	0	0	25.6	0.201	30	No

รูปที่ 4.4 ตัวอย่างข้อมูลโรคเบาหวาน

1	0	1	0	1	0	1	0	1	0	1	0	0.0013	0.024	0.087	0.109	0.08 high
1	0	1	0	1	0	1	0	1	0	1	0	1.00E-04	0.029	0.124	0.128	0.097 high
1	0	1	0	1	0	1	0	1	0	1	0	0.011	0.008	0.073	0.074	0.098 med
1	0	1	0	1	0	1	0	1	0	0	1	1.00E-04	0.023	0.098	0.085	0.115 high
1	0	1	0	1	0	1	0	1	0	1	0	8.00E-04	0.023	0.094	0.099	0.09475 high

รูปที่ 4.5 ตัวอย่างข้อมูลโรคไทรอยด์

Age	Sex
70	1
67	0
57	1

Age	Sex=F	Sex=M
70	1	0
67	0	1
57	1	0

รูปที่ 4.6 การแปลงข้อมูลให้อยู่ในรูป Dummy Variable

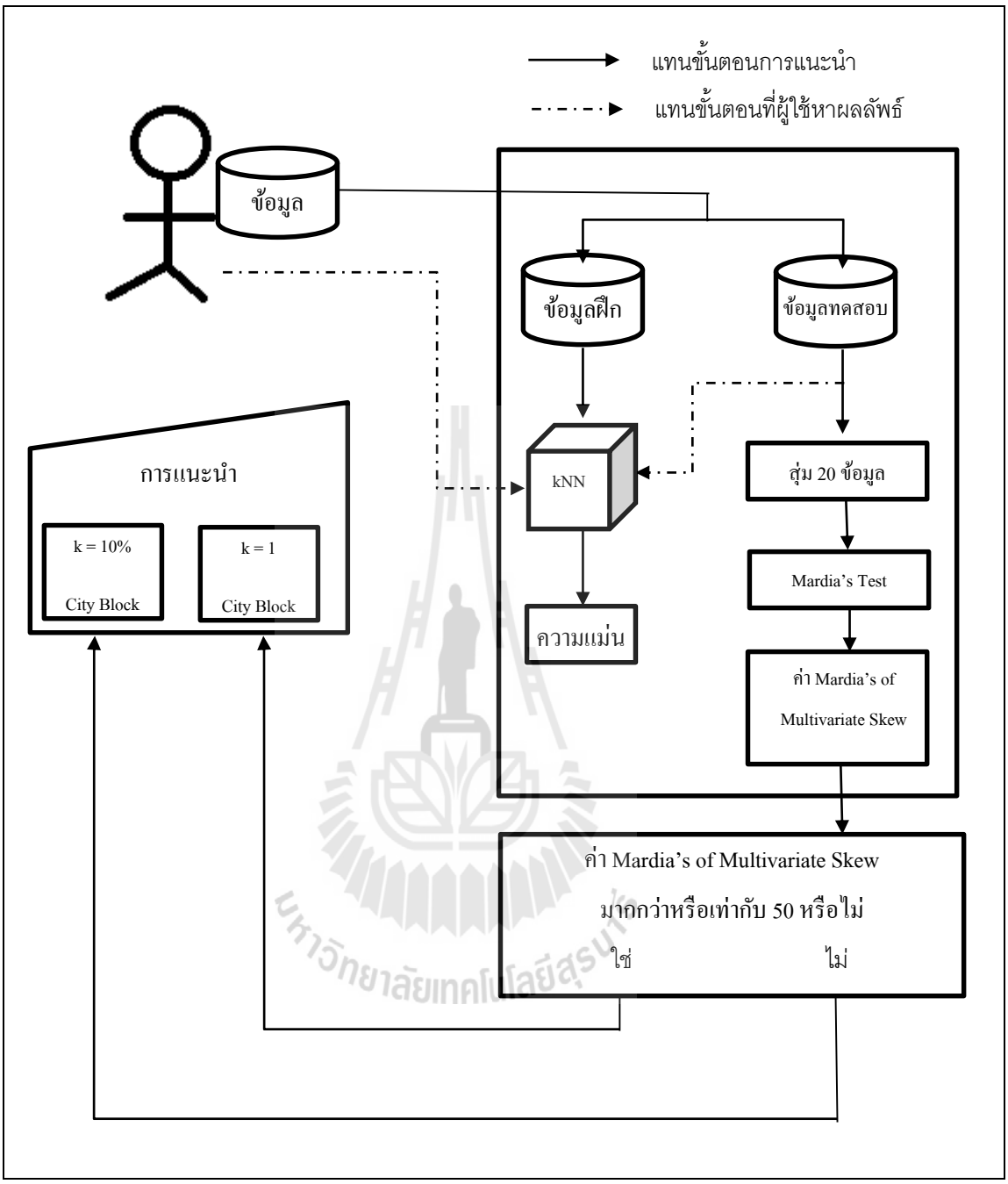
4.2 วิธีการทดสอบประสิทธิภาพ

ในการทดลองจะใช้ข้อมูลทั้ง 5 ข้อมูลคือ ข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคเบาหวาน ข้อมูลโรคหอบหืด และข้อมูลโรคไทรอยด์ ซึ่งการทดลองด้วยอัลกอริทึมจากงานวิจัยนี้ จะมีการวัดการกระจายข้อมูลแบบมาร์เดียเข้าร่วม ซึ่งการใช้โดยทั่วไปจำเป็นต้องมีการวัดประสิทธิภาพร่วมด้วย โดยในขั้นแรกจะเป็นการแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลฝึก และข้อมูลทดสอบซึ่งมีสัดส่วนเป็นร้อยละ 70 ต่อร้อยละ 30 ตามลำดับ หลังจากนั้นจะทำการสุ่มตัวอย่างจากชุดข้อมูลทดสอบ จำนวน 3 ชุด ชุดละ 20 ข้อมูลเพื่อทำการวัดการกระจายโดยมาร์เดีย ซึ่งจะใช้ค่า

Mardia’s of Multivariate Skew ทั้ง 3 ชุดข้อมูลที่สุ่มมาเฉลี่ยกันเพื่อเป็นตัวแทนในการช่วยพิจารณาการตัดสินใจในการกำหนดค่า เค และมาตรวัดระยะทางจากอัลกอริทึมของงานวิจัยนี้ นอกจากนี้สามารถตรวจสอบผลได้จากการสร้างรูปแบบจำแนกโดยการใช้ข้อมูลฝึก ในการสร้างรูปแบบการจำแนกใช้เคเนียร์เรสเนเบอร์ และใช้ข้อมูลทดสอบ ในการทดสอบเพื่อวัดค่าความแม่นได้ ซึ่งกระบวนการจากที่กล่าวมาข้างต้นจะมีลักษณะการทำงานดังรูปที่ 4.7

การทดสอบประสิทธิภาพวิธีการเลือกค่า เค ที่เสนอในงานวิจัยนี้ทดสอบได้โดยการใช้ค่าความแม่นเป็นตัวชี้วัด ซึ่งมาตรวัดระยะทางใดที่ให้ค่าความแม่นสูงถือว่ามาตรวัดนั้นมีประสิทธิภาพในการจำแนกด้วยเคเนียร์เรสเนเบอร์และค่าความแม่นจะบ่งบอกได้ว่ามาตรวัดใดเหมาะสมกับการใช้งาน

มหาวิทยาลัยเทคโนโลยีสุรนารี



รูปที่ 4.7 กระบวนการพิจารณาค่าที่เกี่ยวข้อง

4.3 ผลการทดลองวิธีการจำแนกด้วยอัลกอริทึมจากงานวิจัยนี้

เมื่อทำการทดลองโดยใช้เคเนียร์เรสเนเบอร์ และทำการวัดประสิทธิภาพเพื่อตรวจหาค่าความแม่นที่สูง จากข้อมูลทางการแพทย์ทั้ง 5 ข้อมูล คือ ข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคหอบหืด และข้อมูลโรคไทรอยด์ แสดงดังตารางที่ 4.5 ถึง 4.9

ตารางที่ 4.5 ผลการทดลองกับข้อมูลโรคหัวใจ

D \ k	1	8	16	24	32	41	49	57	65	73	81
Euclidean	0.65	0.57	0.59	0.58	0.57	0.56	0.6	0.62	0.59	0.59	0.63
City Block	**0.74**	0.64	0.63	0.62	0.64	0.64	0.68	0.67	0.67	0.67	0.67
Cosine	0.6	0.58	0.63	0.63	0.6	0.67	0.65	0.69	0.69	0.65	0.64
Correlation	0.59	0.58	0.6	0.63	0.59	0.67	0.65	0.7	0.69	0.65	0.64

ตารางที่ 4.6 ผลการทดลองกับข้อมูลโรคมะเร็งเต้านม

D \ k	1	21	42	63	84	106	127	148	169	190	211
Euclidean	0.98	0.97	0.97	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95
City Block	**0.99**	0.96	0.95	0.94	0.95	0.95	0.95	0.94	0.94	0.94	0.92
Cosine	0.88	0.88	0.86	0.86	0.87	0.85	0.84	0.84	0.82	0.82	0.81
Correlation	0.91	0.91	0.88	0.88	0.88	0.84	0.82	0.82	0.81	0.8	0.8

ตารางที่ 4.7 ผลการทดลองกับข้อมูลโรคเบาหวาน

D \ k	1	23	46	69	92	116	139	162	185	208	231
Euclidean	0.73	0.74	0.74	0.74	0.71	0.7	0.69	0.65	0.65	0.65	0.65
City Block	0.73	**0.77**	0.76	0.75	0.74	0.7	0.68	0.65	0.65	0.65	0.65
Cosine	0.57	0.65	0.66	0.66	0.67	0.66	0.65	0.65	0.65	0.65	0.65
Correlation	0.56	0.68	0.66	0.65	0.68	0.66	0.65	0.65	0.65	0.65	0.65

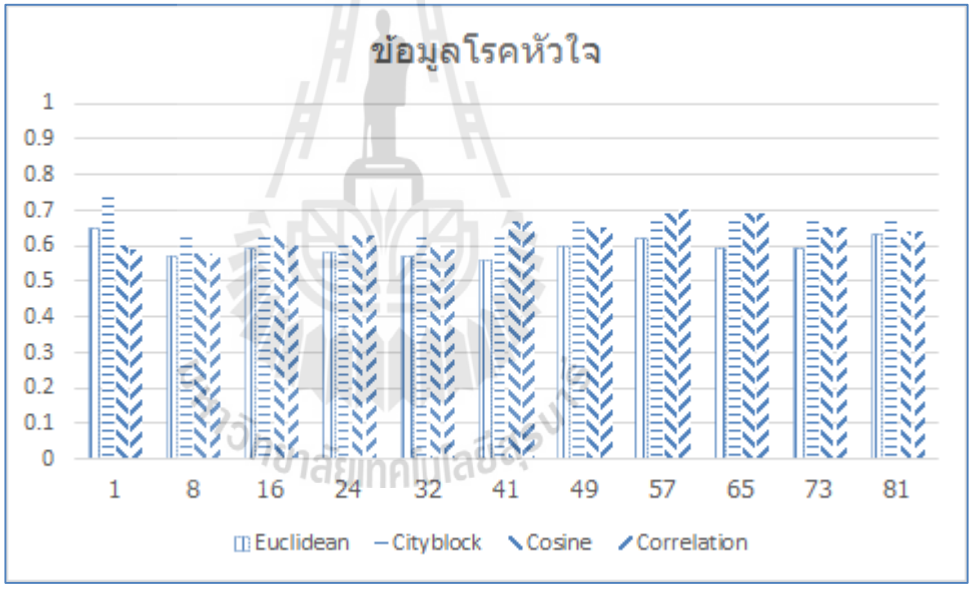
ตารางที่ 4.8 ผลการทดลองกับข้อมูลโรคหอบหืด

D \ k	1	21	41	61	82	102	122	143	163	184	204
Euclidean	0.73	0.80	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
City Block	0.72	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Cosine	0.76	0.81	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Correlation	0.72	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81

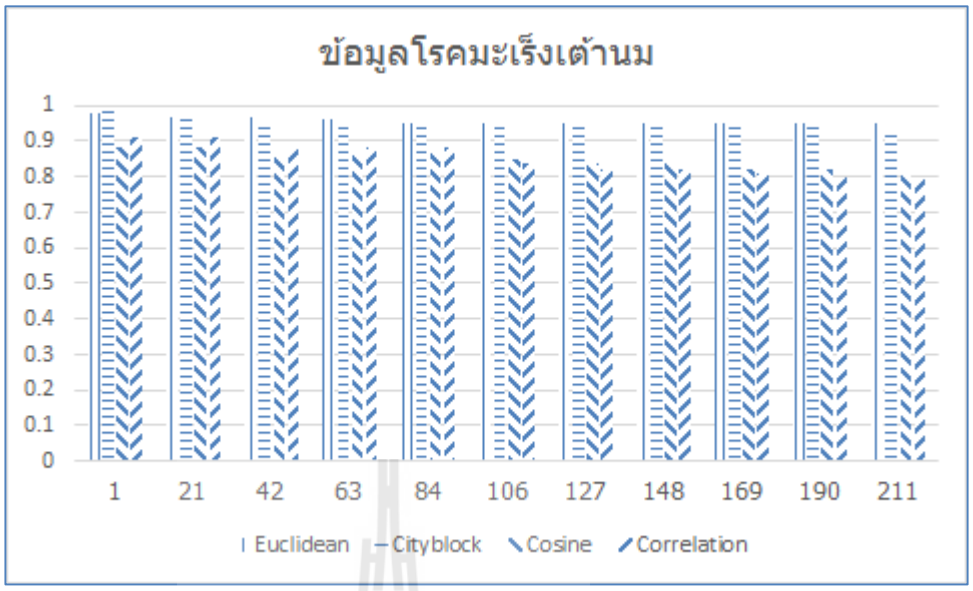
ตารางที่ 4.9 ผลการทดลองกับข้อมูลโรคไทรอยด์

D \ k	1	216	432	648	864	1081	1297	1513	1729	1945	2161
Euclidean	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
City Block	**0.93**	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
Cosine	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
Correlation	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93

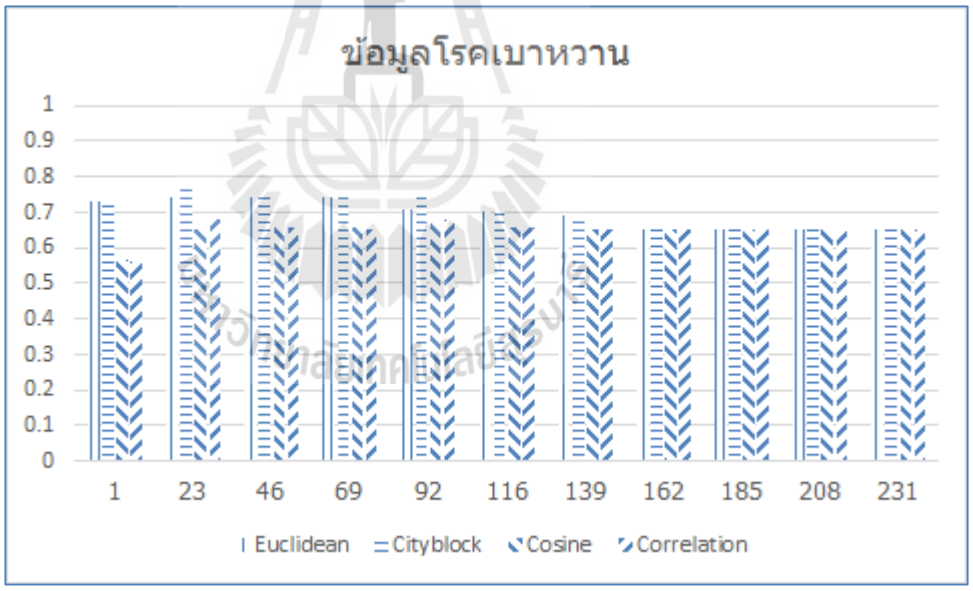
ผลการทดลองจากตารางที่ 4.5 ถึง 4.9 สามารถแสดงเป็นกราฟได้ ดังรูปที่ 4.8 ถึง 4.12 ซึ่งในแนวแกนตั้งคือค่าความแม่น และในแนวแกนนอนคือค่า เค ของเคเนียร์เรสเนเบอร์ และแท่งกราฟแต่ละแท่งบ่งบอกถึงมาตรวัดระยะทางแต่ละมาตรวัด



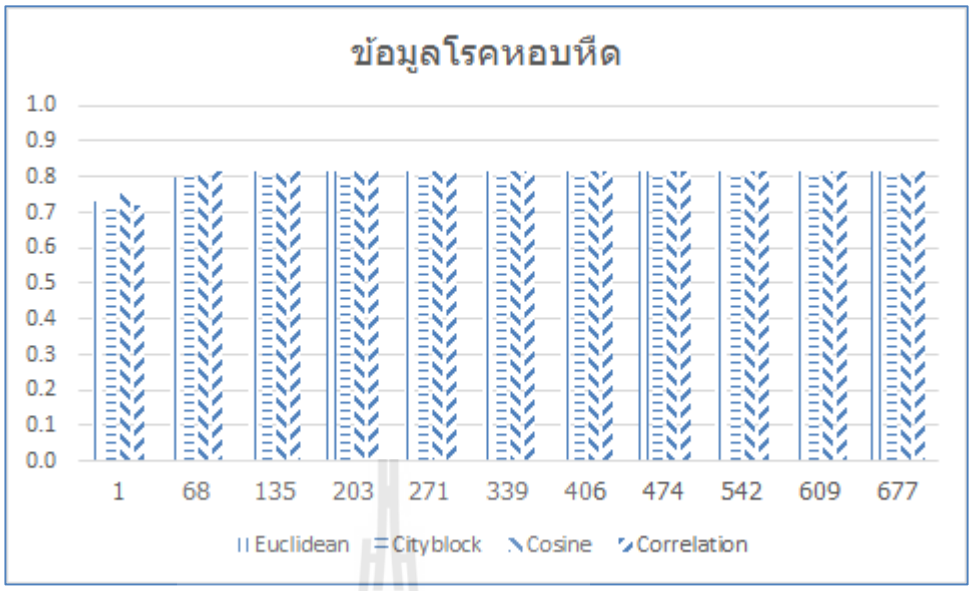
รูปที่ 4.8 ผลการทดลองข้อมูลโรคหัวใจ



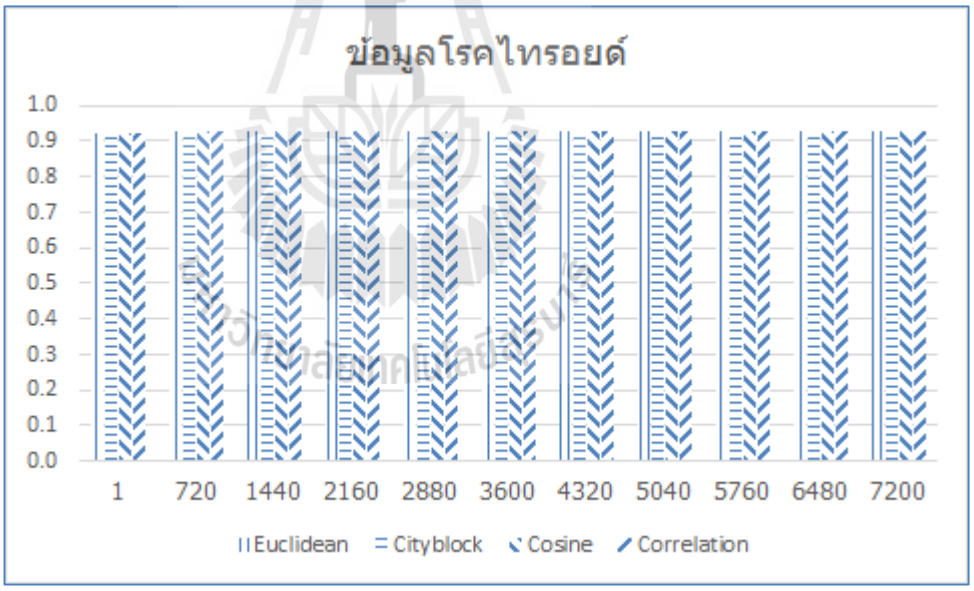
รูปที่ 4.9 ผลการทดลองข้อมูลโรคมะเร็งเต้านม



รูปที่ 4.10 ผลการทดลองข้อมูลโรคเบาหวาน



รูปที่ 4.11 ผลการทดลองข้อมูลโรคหอบหืด



รูปที่ 4.12 ผลการทดลองข้อมูลโรคไทรอยด์

จากตารางที่ 4.5 ถึง 4.9 การหาผลลัพธ์เพื่อให้ได้ซึ่งค่าความแม่นที่สูง จำเป็นต้องเพิ่มค่า เค ขึ้นในแต่ละครั้งเพื่อวัดค่าความแม่นในค่า เค นั้น ๆ จึงทำให้เกิดการวนรอบเพื่อหาค่า เค ทุก ๆ ครั้ง หากมีการใช้กับข้อมูลจำนวนมากและมีการปรับค่า เค มากขึ้น การประมวลของเคเนียร์เรสเนเบอร์ก็จะใช้เวลานานเพื่อให้ได้ผลลัพธ์ในค่า เค นั้น นอกจากนี้เมื่อพบค่า เค ที่ให้ค่าความแม่นสูงสุดได้จาก

ผลทดลองรอบก่อนหน้า แต่ก็ไม่สามารถทราบว่าค่า เค ถัดไปว่าจะให้ค่าความแม่นที่สูงกว่าเดิมหรือไม่ หากไม่ทำการทดสอบประสิทธิภาพต่อ

จากปัญหาข้างต้นสามารถแก้ปัญหาได้โดยการใช้อัลกอริทึมจากงานวิจัยนี้ โดยใช้ผลลัพธ์จากการทดสอบมาร์เดียกับข้อมูลทั้ง 5 ชุด แสดงดังตารางที่ 4.10 เมื่อพิจารณาที่ค่า Mardia's of Multivariate Skew ตามอัลกอริทึมจากงานวิจัยนี้ คือถ้าค่า Mardia's of Multivariate Skew มีค่ามากกว่าหรือเท่ากับ 50 จะแนะนำให้ใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Blockในทางกลับกันหากมีค่าน้อยกว่าจะแนะนำให้ใช้ค่า เค เป็น 10 เปอร์เซ็นต์ของจำนวนข้อมูล

ตารางที่ 4.10 ผลการทดสอบมาร์เดีย

สุ่ม / ข้อมูล	HD	BC	PM	AM	TR
ครั้งที่ 1	100.62	58.37	41.39	49.93	111.29
ครั้งที่ 2	100.84	63.61	38.95	47.33	101.29
ครั้งที่ 3	100.39	71.86	42.12	49.13	180.35
เฉลี่ย	100.62	64.61	40.82	48.80	130.98

จากตารางที่ 4.10 ผลการทดลองในตารางคือค่าของ Mardia's of Multivariate Skew และหมายถึงของ HD, BC, PM, AM, TR คือข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคเบาหวาน ข้อมูลหอบหืด และข้อมูลโรคไทรอยด์ ตามลำดับ เมื่อพิจารณาตามค่าเฉลี่ยของ Mardia's of Multivariate Skew จากการสุ่มข้อมูล 3 ครั้ง สามารถเลือกใช้ค่า เค และมาตรวัดระยะทางตามอัลกอริทึมจากงานวิจัยนี้ได้ดังนี้

ข้อมูลโรคหัวใจ เลือกใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Block

ข้อมูลโรคมะเร็งเต้านม เลือกใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Block

ข้อมูลโรคเบาหวาน เลือกใช้ค่า เค เป็น 10 เปอร์เซ็นต์จากจำนวนข้อมูล กับมาตรวัดระยะทาง City Block

ข้อมูลโรคหอบหืด เลือกใช้ค่า เค เป็น 10 เปอร์เซ็นต์จากจำนวนข้อมูล กับมาตรวัดระยะทาง City Block

ข้อมูลโรคไทรอยด์ เลือกใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Block

เมื่อเลือกใช้ค่า เค และมาตรวัดระยะทางตามอัลกอริทึมจากงานวิจัยนี้ กับข้อมูลทั้ง 5 ข้อมูล ผลลัพธ์ที่ได้ตรงกับตารางที่ 4.5 ถึง 4.9 ซึ่งได้ค่าที่สูงสุดตรงกับในตาราง โดยค่าความแม่นมีลักษณะ

เป็นตัวหนาในตาราง และค่าความแม่นที่ขีดเส้นใต้หมายถึงค่าความแม่นที่สูงที่สุดในมาตรวัดระยะทางนั้น ๆ

4.4 อภิปรายผล

เมื่อใช้อัลกอริทึมจากงานวิจัยนี้กับข้อมูลทั้ง 5 ข้อมูล คือข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคเบาหวาน ข้อมูลโรคหอบหืด และข้อมูลโรคไทรอยด์ พบว่าทั้ง 5 ข้อมูลเมื่อพิจารณาค่าเฉลี่ย Mardia's of Multivariate Skew ตามอัลกอริทึมจากงานวิจัยนี้ จะให้ผลลัพธ์ดังต่อไปนี้

ข้อมูลโรคหัวใจ เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ร้อยละ 74 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคมะเร็งเต้านม เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 99 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคเบาหวาน เมื่อใช้ค่า เค เป็น 10% จากจำนวนข้อมูล และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 77 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคหอบหืด เมื่อใช้ค่า เค เป็น 10% จากจำนวนข้อมูล และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 81 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคไทรอยด์ เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 93 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

เมื่อต้องการค่าความแม่นที่สูงจากการจำแนกโดยใช้เคเนียร์เรสเนเบอร์กับข้อมูลทั้ง 5 ข้อมูล อาจจำเป็นต้องมีการวนรอบการหาค่าความแม่นจากการปรับค่า เค ของเคเนียร์เรสเนเบอร์จนกระทั่งได้ค่าความแม่นที่สูง ซึ่งการปรับค่า เค ในแต่ละรอบเพื่อให้ได้ซึ่งค่าความแม่น การประมวลผลของเคเนียร์เรสเนเบอร์จะค่อนข้างใช้เวลาในการหาผลที่ดีที่สุด ตามจำนวนค่า เค ที่กำหนด แต่หากใช้อัลกอริทึมจากงานวิจัยนี้ จะได้รับคำแนะนำเกี่ยวกับการกำหนดค่า เค และมาตรวัดระยะทาง ซึ่งลดเวลาในกระบวนการการปรับค่า เค เพื่อให้ได้มาซึ่งค่าความแม่นที่สูง และอัลกอริทึมจากงานวิจัยนี้ยังให้ผลลัพธ์ที่ดีตรงกับการวนปรับค่า เค กับข้อมูลทั้ง 5 ข้อมูลด้วย ดังการพิจารณาตามค่าเฉลี่ย Mardia's of Multivariate Skew ทั้งนี้ทั้งนั้นการแนะนำโดยใช้อัลกอริทึมจากงานวิจัยนี้อาจไม่ใช่วิธีการที่ดีสุด แต่เป็นเพียงการเพื่อทางเลือกเพื่อช่วยในการตัดสินของผู้ที่สนใจหรือต้องการใช้งานอัลกอริทึมจากงานวิจัยนี้เพื่อลดเวลาในการกำหนดค่า เค เพื่อให้ได้ค่าความแม่นที่ดีเท่านั้น

บทที่ 4
การทดสอบและอภิปรายผล

ในบทนี้จะกล่าวถึงการทดลองผลจากอัลกอริทึม และข้อมูลต่าง ๆ ที่ใช้ในการทดสอบ ซึ่งจะมีการแสดงผลการทดลองระหว่างการใช้อัลกอริทึมและไม่ใช้อัลกอริทึมจากงานวิจัยนี้ โดยลักษณะการนำเสนอผลการทดลองและกระบวนการดำเนินงานจะมีหลากหลายลักษณะ เช่น กราฟ ตาราง รูปภาพ เป็นต้น

4.1 ข้อมูลที่ใช้ในการทดสอบ

ในการทดสอบแนวทางการเลือกค่า เค ที่เสนอในงานวิจัยนี้ใช้ข้อมูลทางการแพทย์ ซึ่งข้อมูลที่ใช้เหล่านี้เป็นข้อมูลที่มี 2 คลาสเป้าหมาย หรือ 3 คลาสเป้าหมายในการทดสอบเท่านั้น เนื่องจากข้อมูลทางการวินิจฉัยโรคมักจะมีจำนวนคลาสไม่มากโดยประกอบด้วยข้อมูล โรคมะเร็งเต้านม โรคหัวใจ โรคเบาหวาน และโรคไทรอยด์ จาก UCI Machine Learning Repository และข้อมูลโรคหอบหืด จากโรงพยาบาลมหาราช จังหวัดนครราชสีมา โดยทำการรวบรวมข้อมูล ณ วันที่ 8 พฤศจิกายน พ.ศ. 2557 ซึ่งข้อมูลข้างต้นมีลักษณะดังต่อไปนี้

ตารางที่ 4.1 ข้อมูลโรคมะเร็งเต้านม

ลำดับ	คอลัมน์	ค่า
1	Sample Code Number	หมายเลขผู้ป่วย
2	Clump Thickness	1-10
3	Uniformity of Cell Size	1-10
4	Uniformity of Cell Shape	1-10
5	Marginal Adhesion	1-10
6	Single Epithelial Cell Size	1-10
7	Bare Nuclei	1-10
8	Bland Chromatin	1-10
9	Normal Nucleoli	1-10

ตารางที่ 4.1 ข้อมูลโรคมะเร็งเต้านม (ต่อ)

ลำดับ	คอลัมน์	ค่า
10	Mitoses	1-10
11	Class	Benign, Malignant

จากตารางที่ 4.1 เป็นข้อมูลโรคมะเร็งเต้านม ซึ่งประกอบด้วย 11 คอลัมน์ จำนวนข้อมูลทั้งหมด 699 ข้อมูล ซึ่งในคอลัมน์ Sample Code Number จะไม่ถูกทำมาวิเคราะห์เนื่องจากเป็นเพียงหมายเลขผู้ป่วยซึ่งไม่ส่งผลกับการจำแนก

ตารางที่ 4.2 ข้อมูลโรคหัวใจ

ลำดับ	คอลัมน์	ค่า
1	Age	29-77
2	Sex	Male, Female
3	Chest Pain Type	0, 1, 2, 3
4	Resting Blood Pressure	94-200
5	Serum Cholestoral	126-564
6	Fasting Blood Sugar > 120 mg/dl	0, 1
7	Resting Electrocardiographic Results	0, 1
8	Maximum Heart Rate Achieved	71-202
9	Exercise Induced Angina	0, 1
10	Oldpeak	0-6.2
11	The Slope of The Peak Exercise ST Segment	0, 1, 2
12	Number of Major Vessels	0, 1, 2, 3
13	Thal	0, 1, 2
14	Class	Absence, Presence

จากตารางที่ 4.2 เป็นข้อมูลโรคหัวใจ ซึ่งประกอบด้วย 14 คอลัมน์ จำนวนข้อมูลทั้งหมด 270 ข้อมูล

ตารางที่ 4.3 ข้อมูลโรคเบาหวาน

ลำดับ	คอลัมน์	ค่า
1	Number of Times Pregnant	0-17
2	Plasma Glucose Concentration a 2 Hours in an Oral Glucose Tolerance Test	0-199
3	Diastolic Blood Pressure	0-122
4	Triceps Skin Fold Thickness	0-99
5	2-Hour Serum Insulin	0-846
6	Body Mass Index	0-67.1
7	Diabetes Pedigree Function	0.078-2.42
8	Age	21-81
9	Class	Yes, No

จากตารางที่ 4.3 เป็นข้อมูลโรคเบาหวาน ซึ่งประกอบด้วย 9 คอลัมน์ จำนวนข้อมูลทั้งหมด 768 ข้อมูล

ตารางที่ 4.4 ข้อมูลโรคหอบหืด

ลำดับ	คอลัมน์	ค่า
1	Case Number	หมายเลขผู้ป่วย
2	Age Respondents (Years)	35-64
3	Gender of Respondents	0, 1
4	Highest Education Level	1, 2, 3, 4
5	Marital Status	1, 2, 3, 4, 5
6	Religion	1, 2, 3
7	Smoking	0, 1, 2
8	Exercise	0, 1
9	Weight (kg.)	37.2-113.3
10	Weight (cm.)	141-192
11	Waist (cm.)	57-119
12	Percent Body Fat	11.7-47.6

ตารางที่ 4.4 ข้อมูลโรคหอบหืด (ต่อ)

ลำดับ	คอลัมน์	ค่า
13	PA level	Low, Moderate, High

จากตารางที่ 4.4 เป็นข้อมูลโรคหอบหืด ซึ่งประกอบด้วย 13 คอลัมน์ จำนวนข้อมูลทั้งหมด 677 ข้อมูล

ข้อมูลโรคไทรอยด์ จำนวนข้อมูลทั้งหมด 7200 ข้อมูล โดยมี 22 คอลัมน์ ข้อมูลนี้ได้จากแหล่งข้อมูลมาตรฐาน UCI Machine Learning Repository ซึ่งถูกบริจารในปี ค.ศ. 1987 โดย Peter Turney

จากรูปที่ 4.1 เป็นรูปแสดงตัวอย่างของข้อมูลโรคหอบหืด ซึ่งจะพบว่ามีเพียง 12 คอลัมน์เท่านั้นที่ถูกใช้งาน โดยคอลัมน์ที่ไม่ส่งผลต่อโรคจะถูกตัดทิ้ง เช่น คอลัมน์หมายเลขผู้ป่วย หมายเลขประจำตัวประชาชน หมายเลขโทรศัพท์ เป็นต้น

age	sex	edu	ms	rel	smoking	exercise	wt	ht	wc	bodyfat	PA_level
54	1	2	2	1	0	1	62.4	163	75.5	33	moderate
52	1	3	2	1	1	0	47.5	153	65.5	27.6	low
51	1	3	2	1	0	1	57.1	160	83	31.6	moderate
55	1	3	1	1	0	0	57.7	150	87	38.2	low
57	1	2	2	1	0	0	53.4	158	76	32.2	low

รูปที่ 4.1 ตัวอย่างข้อมูลโรคหอบหืด

ข้อมูลที่ใช้ในการทดสอบจะมีลักษณะเป็นตัวเลข และมีคลาสเป้าหมายเป็นตัวอักษร โดยมีลักษณะข้อมูลตัวอย่าง ดังรูปที่ 4.1 ถึง 4.5 ซึ่งเป็นบางส่วนของข้อมูล ทั้งนี้ในบางข้อมูลจำเป็นต้องมีการจัดการข้อมูลก่อน การดำเนินการทดสอบเช่นข้อมูลที่มีลักษณะที่วัดไม่ได้ เช่น เพศชาย เพศหญิง จึงจำเป็นต้องแปลงข้อมูลนั้น ๆ ให้อยู่ในรูปของ Dummy Variable เพื่อช่วยให้อัลกอริทึมสามารถใช้งานกับข้อมูลได้ โดยที่ลักษณะการแปลงข้อมูลให้อยู่ในรูป Dummy Variable มีลักษณะดังรูปที่ 4.6

130	322	0	1	1	0	109	0	1	2.4	0	1	0	3	0	1	0 sick
115	564	0	1	1	0	160	0	1	1.6	0	1	0	0	0	0	1 healthy
124	261	0	1	0	1	141	0	1	0.3	0	0	1	0	0	0	1 sick
128	263	0	1	0	1	105	1	0	0.2	0	1	0	1	0	0	1 healthy
120	269	0	1	1	0	121	1	0	0.2	0	0	1	1	0	1	0 healthy
120	177	0	1	0	1	140	0	1	0.4	0	0	1	0	0	0	1 healthy

รูปที่ 4.2 ตัวอย่างข้อมูลโรคหัวใจ

6	8	8	1	3	4	3	7	1	Benign
4	1	1	3	2	1	3	1	1	Benign
8	10	10	8	7	10	9	7	1	Malignant
1	1	1	1	2	10	3	1	1	Benign
2	1	2	1	2	1	3	1	1	Benign

รูปที่ 4.3 ตัวอย่างข้อมูลโรคมะเร็งเต้านม

8	183	64	0	0	23.3	0.672	32	Yes
1	89	66	23	94	28.1	0.167	21	No
0	137	40	35	168	43.1	2.288	33	Yes
5	116	74	0	0	25.6	0.201	30	No

รูปที่ 4.4 ตัวอย่างข้อมูลโรคเบาหวาน

1	0	1	0	1	0	1	0	1	0	1	0	0.0013	0.024	0.087	0.109	0.08 high
1	0	1	0	1	0	1	0	1	0	1	0	1.00E-04	0.029	0.124	0.128	0.097 high
1	0	1	0	1	0	1	0	1	0	1	0	0.011	0.008	0.073	0.074	0.098 med
1	0	1	0	1	0	1	0	1	0	0	1	1.00E-04	0.023	0.098	0.085	0.115 high
1	0	1	0	1	0	1	0	1	0	1	0	8.00E-04	0.023	0.094	0.099	0.09475 high

รูปที่ 4.5 ตัวอย่างข้อมูลโรคไทรอยด์

Age	Sex
70	1
67	0
57	1

Age	Sex=F	Sex=M
70	1	0
67	0	1
57	1	0

รูปที่ 4.6 การแปลงข้อมูลให้อยู่ในรูป Dummy Variable

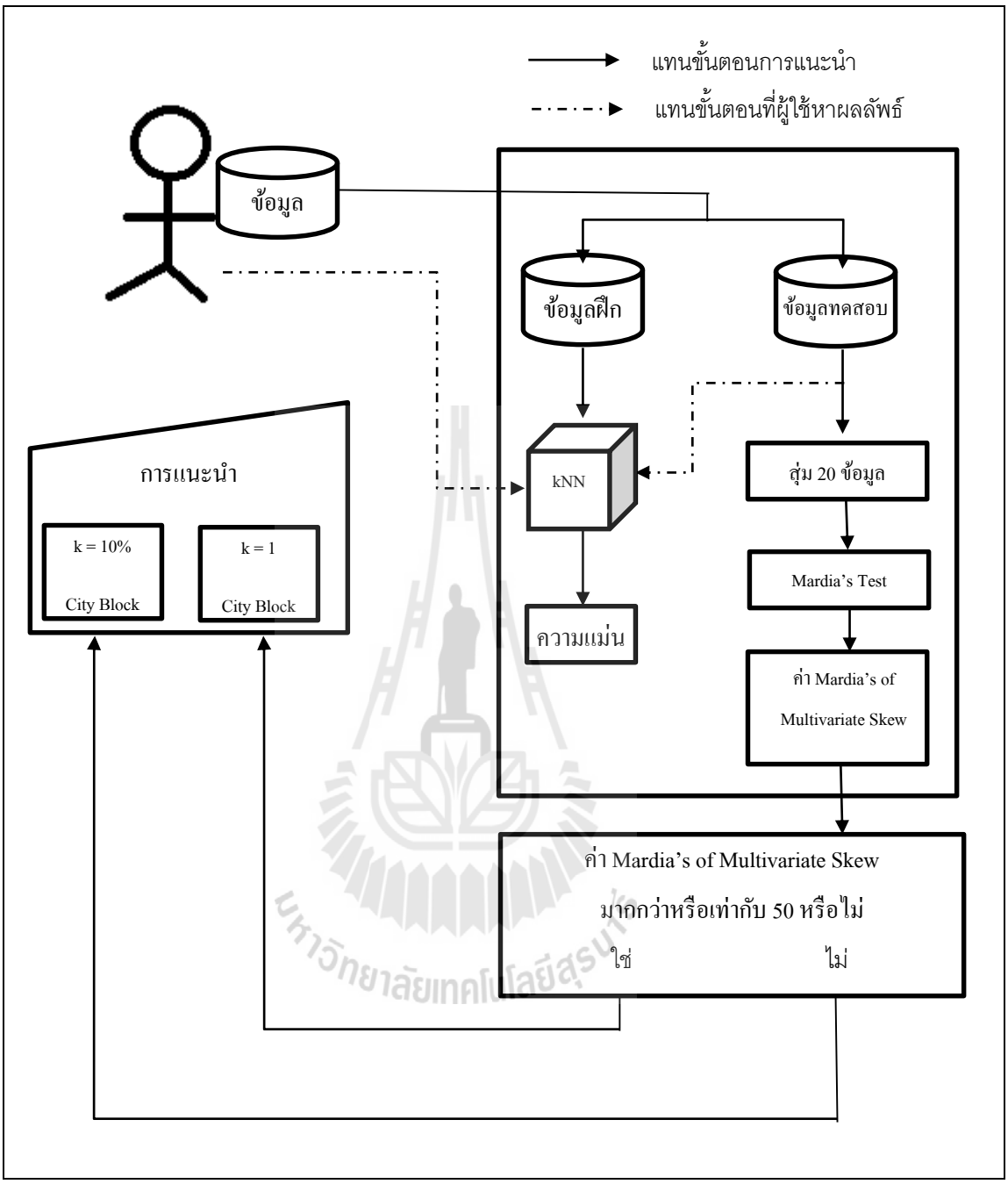
4.2 วิธีการทดสอบประสิทธิภาพ

ในการทดลองจะใช้ข้อมูลทั้ง 5 ข้อมูลคือ ข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคเบาหวาน ข้อมูลโรคหอบหืด และข้อมูลโรคไทรอยด์ ซึ่งการทดลองด้วยอัลกอริทึมจากงานวิจัยนี้ จะมีการวัดการกระจายข้อมูลแบบมาร์เดียเข้าร่วม ซึ่งการใช้โดยทั่วไปจำเป็นต้องมีการวัดประสิทธิภาพร่วมด้วย โดยในขั้นแรกจะเป็นการแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลฝึก และข้อมูลทดสอบซึ่งมีสัดส่วนเป็นร้อยละ 70 ต่อร้อยละ 30 ตามลำดับ หลังจากนั้นจะทำการสุ่มตัวอย่างจากชุดข้อมูลทดสอบ จำนวน 3 ชุด ชุดละ 20 ข้อมูลเพื่อทำการวัดการกระจายโดยมาร์เดีย ซึ่งจะใช้ค่า

Mardia’s of Multivariate Skew ทั้ง 3 ชุดข้อมูลที่สุ่มมาเฉลี่ยกันเพื่อเป็นตัวแทนในการช่วยพิจารณาการตัดสินใจในการกำหนดค่า เค และมาตรวัดระยะทางจากอัลกอริทึมของงานวิจัยนี้ นอกจากนี้สามารถตรวจสอบผลได้จากการสร้างรูปแบบจำแนกโดยการใช้ข้อมูลฝึก ในการสร้างรูปแบบการจำแนกใช้เคเนียร์เรสเนเบอร์ และใช้ข้อมูลทดสอบ ในการทดสอบเพื่อวัดค่าความแม่นได้ ซึ่งกระบวนการจากที่กล่าวมาข้างต้นจะมีลักษณะการทำงานดังรูปที่ 4.7

การทดสอบประสิทธิภาพวิธีการเลือกค่า เค ที่เสนอในงานวิจัยนี้ทดสอบได้โดยการใช้ค่าความแม่นเป็นตัวชี้วัด ซึ่งมาตรวัดระยะทางใดที่ให้ค่าความแม่นสูงถือว่ามาตรวัดนั้นมีประสิทธิภาพในการจำแนกด้วยเคเนียร์เรสเนเบอร์และค่าความแม่นจะบ่งบอกได้ว่ามาตรวัดใดเหมาะสมกับการใช้งาน

มหาวิทยาลัยเทคโนโลยีสุรนารี



รูปที่ 4.7 กระบวนการพิจารณาค่าที่เกี่ยวข้อง

4.3 ผลการทดลองวิธีการจำแนกด้วยอัลกอริทึมจากงานวิจัยนี้

เมื่อทำการทดลองโดยใช้เคเนียร์เรสเนเบอร์ และทำการวัดประสิทธิภาพเพื่อตรวจหาค่าความแม่นที่สูง จากข้อมูลทางการแพทย์ทั้ง 5 ข้อมูล คือ ข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคหอบหืด และข้อมูลโรคไทรอยด์ แสดงดังตารางที่ 4.5 ถึง 4.9

ตารางที่ 4.5 ผลการทดลองกับข้อมูลโรคหัวใจ

D \ k	1	8	16	24	32	41	49	57	65	73	81
Euclidean	0.65	0.57	0.59	0.58	0.57	0.56	0.6	0.62	0.59	0.59	0.63
City Block	**0.74**	0.64	0.63	0.62	0.64	0.64	0.68	0.67	0.67	0.67	0.67
Cosine	0.6	0.58	0.63	0.63	0.6	0.67	0.65	0.69	0.69	0.65	0.64
Correlation	0.59	0.58	0.6	0.63	0.59	0.67	0.65	0.7	0.69	0.65	0.64

ตารางที่ 4.6 ผลการทดลองกับข้อมูลโรคมะเร็งเต้านม

D \ k	1	21	42	63	84	106	127	148	169	190	211
Euclidean	0.98	0.97	0.97	0.96	0.95	0.95	0.95	0.95	0.95	0.95	0.95
City Block	**0.99**	0.96	0.95	0.94	0.95	0.95	0.95	0.94	0.94	0.94	0.92
Cosine	0.88	0.88	0.86	0.86	0.87	0.85	0.84	0.84	0.82	0.82	0.81
Correlation	0.91	0.91	0.88	0.88	0.88	0.84	0.82	0.82	0.81	0.8	0.8

ตารางที่ 4.7 ผลการทดลองกับข้อมูลโรคเบาหวาน

D \ k	1	23	46	69	92	116	139	162	185	208	231
Euclidean	0.73	0.74	0.74	0.74	0.71	0.7	0.69	0.65	0.65	0.65	0.65
City Block	0.73	**0.77**	0.76	0.75	0.74	0.7	0.68	0.65	0.65	0.65	0.65
Cosine	0.57	0.65	0.66	0.66	0.67	0.66	0.65	0.65	0.65	0.65	0.65
Correlation	0.56	0.68	0.66	0.65	0.68	0.66	0.65	0.65	0.65	0.65	0.65

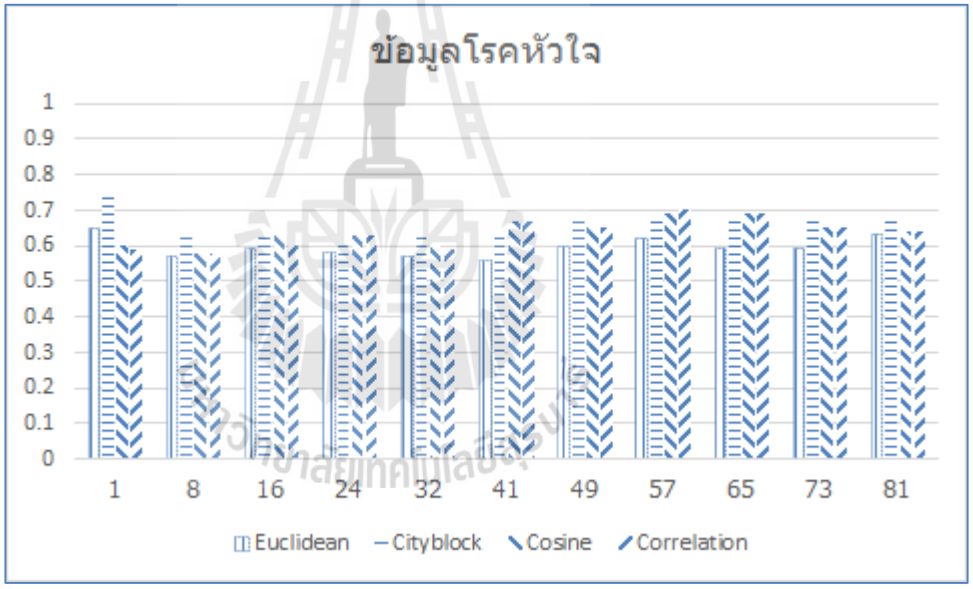
ตารางที่ 4.8 ผลการทดลองกับข้อมูลโรคหอบหืด

D \ k	1	21	41	61	82	102	122	143	163	184	204
Euclidean	0.73	0.80	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
City Block	0.72	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Cosine	0.76	0.81	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Correlation	0.72	**0.81**	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81

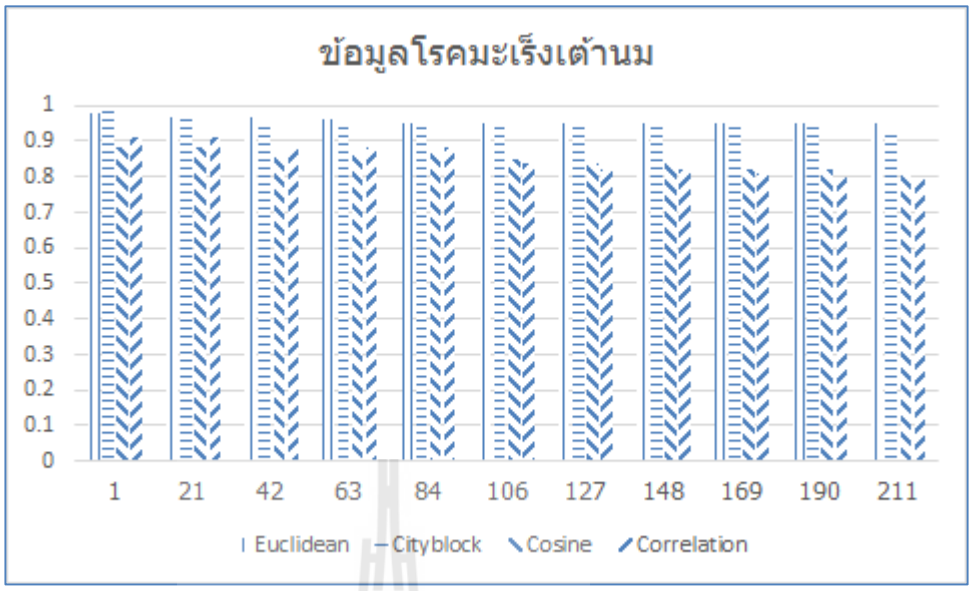
ตารางที่ 4.9 ผลการทดลองกับข้อมูลโรคไทรอยด์

D \ k	1	216	432	648	864	1081	1297	1513	1729	1945	2161
Euclidean	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
City Block	**0.93**	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
Cosine	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93
Correlation	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93	0.93

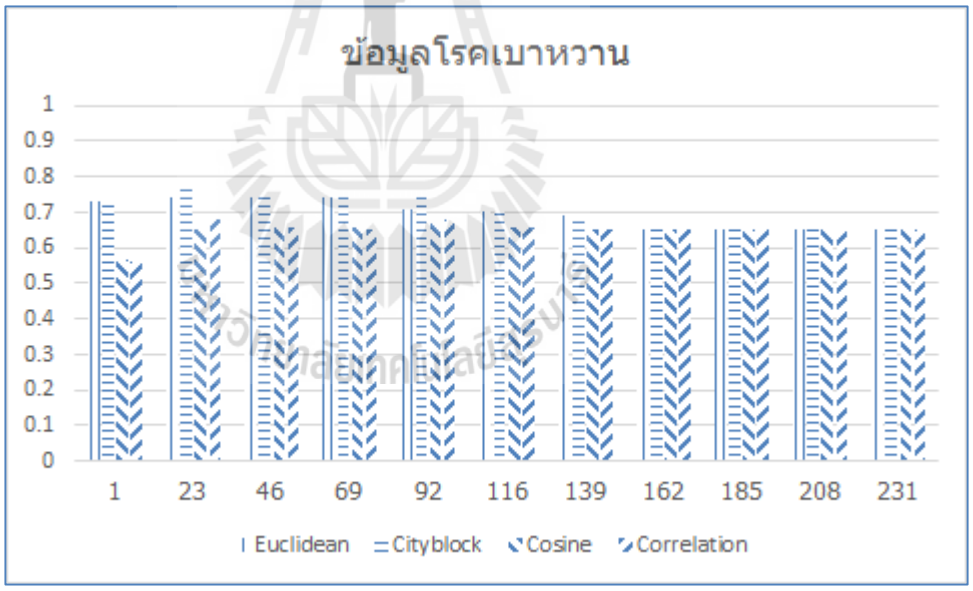
ผลการทดลองจากตารางที่ 4.5 ถึง 4.9 สามารถแสดงเป็นกราฟได้ ดังรูปที่ 4.8 ถึง 4.12 ซึ่งในแนวแกนตั้งคือค่าความแม่น และในแนวแกนนอนคือค่า เค ของเคเนียร์เรสเนเบอร์ และแท่งกราฟแต่ละแท่งบ่งบอกถึงมาตรวัดระยะทางแต่ละมาตรวัด



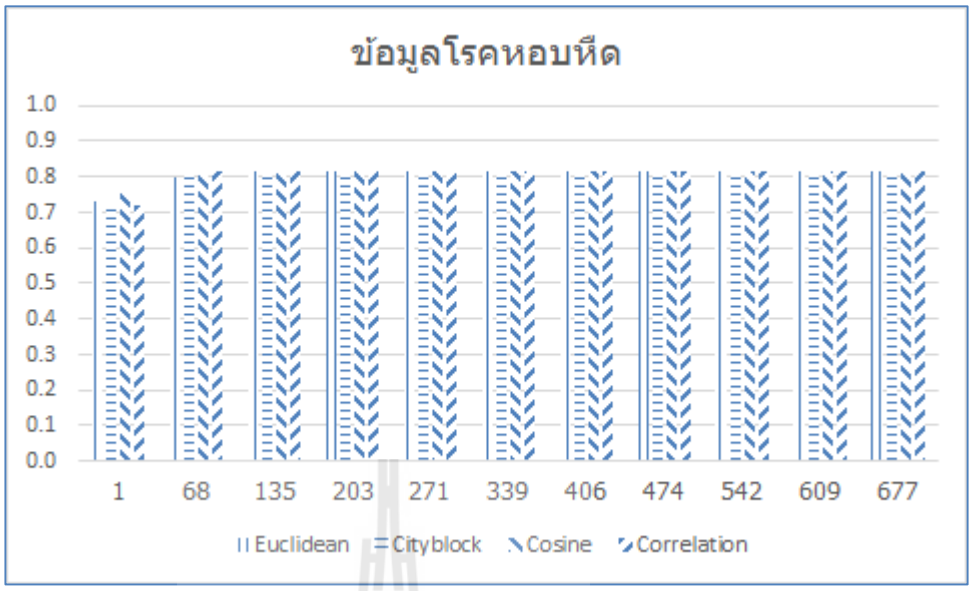
รูปที่ 4.8 ผลการทดลองข้อมูลโรคหัวใจ



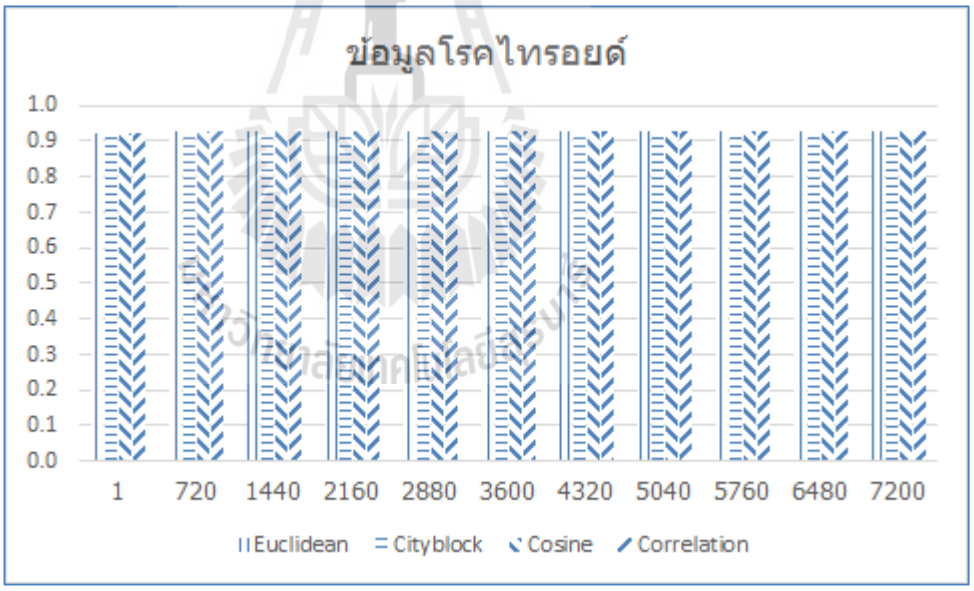
รูปที่ 4.9 ผลการทดลองข้อมูลโรคมะเร็งเต้านม



รูปที่ 4.10 ผลการทดลองข้อมูลโรคเบาหวาน



รูปที่ 4.11 ผลการทดลองข้อมูลโรคหอบหืด



รูปที่ 4.12 ผลการทดลองข้อมูลโรคไทรอยด์

จากตารางที่ 4.5 ถึง 4.9 การหาผลลัพธ์เพื่อให้ได้ซึ่งค่าความแม่นที่สูง จำเป็นต้องเพิ่มค่า เค ขึ้นในแต่ละครั้งเพื่อวัดค่าความแม่นในค่า เค นั้น ๆ จึงทำให้เกิดการวนรอบเพื่อหาค่า เค ทุก ๆ ครั้ง หากมีการใช้กับข้อมูลจำนวนมากและมีการปรับค่า เค มากขึ้น การประมวลของเคเนียร์เรสเนเบอร์ก็จะใช้เวลานานเพื่อให้ได้ผลลัพธ์ในค่า เค นั้น นอกจากนี้เมื่อพบค่า เค ที่ให้ค่าความแม่นสูงสุดได้จาก

ผลทดลองรอบก่อนหน้า แต่ก็ไม่สามารถทราบว่าค่า เค ถัดไปว่าจะให้ค่าความแม่นที่สูงกว่าเดิมหรือไม่ หากไม่ทำการทดสอบประสิทธิภาพต่อ

จากปัญหาข้างต้นสามารถแก้ปัญหาได้โดยการใช้อัลกอริทึมจากงานวิจัยนี้ โดยใช้ผลลัพธ์จากการทดสอบมาร์เดียกับข้อมูลทั้ง 5 ชุด แสดงดังตารางที่ 4.10 เมื่อพิจารณาที่ค่า Mardia's of Multivariate Skew ตามอัลกอริทึมจากงานวิจัยนี้ คือถ้าค่า Mardia's of Multivariate Skew มีค่ามากกว่าหรือเท่ากับ 50 จะแนะนำให้ใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Blockในทางกลับกันหากมีค่าน้อยกว่าจะแนะนำให้ใช้ค่า เค เป็น 10 เปอร์เซ็นต์ของจำนวนข้อมูล

ตารางที่ 4.10 ผลการทดสอบมาร์เดีย

สุ่ม / ข้อมูล	HD	BC	PM	AM	TR
ครั้งที่ 1	100.62	58.37	41.39	49.93	111.29
ครั้งที่ 2	100.84	63.61	38.95	47.33	101.29
ครั้งที่ 3	100.39	71.86	42.12	49.13	180.35
เฉลี่ย	100.62	64.61	40.82	48.80	130.98

จากตารางที่ 4.10 ผลการทดลองในตารางคือค่าของ Mardia's of Multivariate Skew และหมายถึงของ HD, BC, PM, AM, TR คือข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคเบาหวาน ข้อมูลหอบหืด และข้อมูลโรคไทรอยด์ ตามลำดับ เมื่อพิจารณาตามค่าเฉลี่ยของ Mardia's of Multivariate Skew จากการสุ่มข้อมูล 3 ครั้ง สามารถเลือกใช้ค่า เค และมาตรวัดระยะทางตามอัลกอริทึมจากงานวิจัยนี้ได้ดังนี้

ข้อมูลโรคหัวใจ เลือกใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Block

ข้อมูลโรคมะเร็งเต้านม เลือกใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Block

ข้อมูลโรคเบาหวาน เลือกใช้ค่า เค เป็น 10 เปอร์เซ็นต์จากจำนวนข้อมูล กับมาตรวัดระยะทาง City Block

ข้อมูลโรคหอบหืด เลือกใช้ค่า เค เป็น 10 เปอร์เซ็นต์จากจำนวนข้อมูล กับมาตรวัดระยะทาง City Block

ข้อมูลโรคไทรอยด์ เลือกใช้ค่า เค เป็น 1 กับมาตรวัดระยะทาง City Block

เมื่อเลือกใช้ค่า เค และมาตรวัดระยะทางตามอัลกอริทึมจากงานวิจัยนี้ กับข้อมูลทั้ง 5 ข้อมูล ผลลัพธ์ที่ได้ตรงกับตารางที่ 4.5 ถึง 4.9 ซึ่งได้ค่าที่สูงสุดตรงกับในตาราง โดยค่าความแม่นมีลักษณะ

เป็นตัวหนาในตาราง และค่าความแม่นที่ขีดเส้นใต้หมายถึงค่าความแม่นที่สูงที่สุดในมาตรวัดระยะทางนั้น ๆ

4.4 อภิปรายผล

เมื่อใช้อัลกอริทึมจากงานวิจัยนี้กับข้อมูลทั้ง 5 ข้อมูล คือข้อมูลโรคหัวใจ ข้อมูลโรคมะเร็งเต้านม ข้อมูลโรคเบาหวาน ข้อมูลโรคหอบหืด และข้อมูลโรคไทรอยด์ พบว่าทั้ง 5 ข้อมูลเมื่อพิจารณาค่าเฉลี่ย Mardia's of Multivariate Skew ตามอัลกอริทึมจากงานวิจัยนี้ จะให้ผลลัพธ์ดังต่อไปนี้

ข้อมูลโรคหัวใจ เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ร้อยละ 74 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคมะเร็งเต้านม เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 99 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคเบาหวาน เมื่อใช้ค่า เค เป็น 10% จากจำนวนข้อมูล และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 77 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคหอบหืด เมื่อใช้ค่า เค เป็น 10% จากจำนวนข้อมูล และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 81 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

ข้อมูลโรคไทรอยด์ เมื่อใช้ค่า เค เป็น 1 และมาตรวัดระยะทาง City Block จะได้ค่าความแม่นที่ ร้อยละ 93 ซึ่งเป็นค่าที่สูงที่สุดในตาราง

เมื่อต้องการค่าความแม่นที่สูงจากการจำแนกโดยใช้เคเนียร์เรสเนเบอร์กับข้อมูลทั้ง 5 ข้อมูล อาจจำเป็นต้องมีการวนรอบการหาค่าความแม่นจากการปรับค่า เค ของเคเนียร์เรสเนเบอร์จนกระทั่งได้ค่าความแม่นที่สูง ซึ่งการปรับค่า เค ในแต่ละรอบเพื่อให้ได้ซึ่งค่าความแม่น การประมวลผลของเคเนียร์เรสเนเบอร์จะค่อนข้างใช้เวลาในการหาผลที่ดีที่สุด ตามจำนวนค่า เค ที่กำหนด แต่หากใช้อัลกอริทึมจากงานวิจัยนี้ จะได้รับคำแนะนำเกี่ยวกับการกำหนดค่า เค และมาตรวัดระยะทาง ซึ่งลดเวลาในกระบวนการการปรับค่า เค เพื่อให้ได้มาซึ่งค่าความแม่นที่สูง และอัลกอริทึมจากงานวิจัยนี้ยังให้ผลลัพธ์ที่ดีตรงกับการวนปรับค่า เค กับข้อมูลทั้ง 5 ข้อมูลด้วย ดังการพิจารณาตามค่าเฉลี่ย Mardia's of Multivariate Skew ทั้งนี้ทั้งนั้นการแนะนำโดยใช้อัลกอริทึมจากงานวิจัยนี้อาจไม่ใช่วิธีการที่ดีสุด แต่เป็นเพียงการเพื่อทางเลือกเพื่อช่วยในการตัดสินของผู้ที่สนใจหรือต้องการใช้งานอัลกอริทึมจากงานวิจัยนี้เพื่อลดเวลาในการกำหนดค่า เค เพื่อให้ได้ค่าความแม่นที่ดีเท่านั้น

บทที่ 5
สรุปผลงานวิจัยและข้อเสนอแนะ

การใช้งานเคเนียร์เรสเนเบอร์จำเป็นต้องมีการกำหนดค่า เค ก่อนการใช้งานเสมอ และการกำหนดค่า เค มักส่งผลกับค่าความแม่นในการจำแนก ในปัจจุบันจึงมีการหาวิธีเพื่อกำหนดค่า เค ในการใช้งานเคเนียร์เรสเนเบอร์โดยที่ค่า เค นั้นให้ค่าความแม่นที่สูง

ลักษณะการทำงานของเคเนียร์เรสเนเบอร์เป็นลักษณะ Lazy-Learning หรือการเรียนรู้แบบขี้เกียจ คือการทำงานจะเริ่มขึ้นเมื่อมีงานเข้ามาเท่านั้น และใช้การพิจารณาด้วยสภาพแวดล้อมที่คุ้นเคยในการจำแนก ซึ่งเป็นการทำงานในลักษณะที่สอดคล้องกับการวินิจฉัยโรคต่าง ๆ ในทางการแพทย์ โดยใช้ข้อมูลหรือลักษณะอาการเดิมในอดีตเพื่อนำมาซึ่งการวินิจฉัยคนไข้รายใหม่ ว่าป่วยหรือไม่

ในงานวิจัยนี้จึงมีการแนะนำเกี่ยวกับการกำนหดค่า เค และมาตรวัดระยะทาง ซึ่งการแนะนำและการใช้งานสะดวกและง่ายสำหรับการนำไปใช้ โดยพิจารณาจากการทดสอบแบบมาร์เดียกับข้อมูลที่ต้องการ

5.1 ขั้นตอนการดำเนินงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์คือการแนะนำการเลือกใช้ค่า เค ของเคเนียร์เรสเนเบอร์ โดยมีการทดลองกับข้อมูลทางการแพทย์ ซึ่งขั้นตอนโดยสรุปเป็นดังต่อไปนี้

1. ศึกษากระบวนการการทำงานของการจำแนกด้วยวิธีเคเนียร์เรสเนเบอร์
2. ออกแบบอัลกอริทึมโดยใช้การทดสอบแบบมาร์เดียในการพิจารณาการกำหนดค่า เค
3. ทดลองและเปรียบเทียบผลอัลกอริทึมจากงานวิจัยนี้กับการใช้เคเนียร์เรสเนเบอร์ในการหาค่าความแม่นที่สูง จากการปรับค่า เค แต่ละครั้งเพื่อให้ได้ค่าความแม่น
4. ให้คำแนะนำสำหรับผู้ใช้ที่ต้องการกำหนดค่า เค และมาตรวัดระยะทาง เพื่อเพิ่มทางเลือกสำหรับผู้ที่ต้องการความสะดวกรวดเร็ว และให้ค่าความแม่นที่อยู่ในช่วงที่สูง

5.2 สรุปผลงานวิจัย

การใช้งานเคเนียร์เรสเนเบอร์ในการจำแนก จำเป็นต้องมีการกำหนดค่า เค และมาตรวัดระยะทางก่อนการจำแนกเสมอ และเมื่อผู้ใช้งานต้องการค่าความแม่นที่สูง อาจจำเป็นต้องมีการ

กำหนดค่า เค และมีการปรับค่า เค เพื่อให้ได้มาซึ่งค่าความแม่นที่ผู้ใช้ต้องการ และเมื่อมีการปรับค่า เค มากขึ้น การประมวลผลของเคเนียร์เรสเนเบอร์ก็จะใช้เวลานานขึ้นตามค่า เค ที่เพิ่มขึ้น อัลกอริทึมจากงานวิจัยนี้เป็นการแนะนำเกี่ยวกับการกำหนดค่า เค และมาตรวัดระยะทาง ซึ่งช่วยลดเวลาในส่วนของการปรับค่า เค สำหรับผู้ใช้ที่ต้องการค่าความแม่นที่สูง ในงานวิจัยนี้ใช้ข้อมูลทางการแพทย์ในการทดลอง ซึ่งประกอบด้วยข้อมูลดังต่อไปนี้

- ข้อมูลโรคหัวใจ
- ข้อมูลโรคมะเร็งเต้านม
- ข้อมูลโรคเบาหวาน
- ข้อมูลโรคหอบหืด
- ข้อมูลโรคไทรอยด์

ทั้งนี้อัลกอริทึมจากงานวิจัยนี้เป็นการแนะนำในการกำหนดค่าพารามิเตอร์บ้างส่วนของเคเนียร์เรสเนเบอร์ ซึ่งการแนะนำจากอัลกอริทึมนี้อาจไม่ใช่วิธีการที่ดีที่สุด แต่สามารถเป็นทางเลือกให้กับผู้ที่สนใจหรือผู้ที่ต้องการใช้งาน เพื่อลดเวลาในการวิเคราะห์หรือกระบวนการบางอย่างให้แก่ผู้สนใจหรือผู้ที่ต้องการใช้งาน นอกจากนี้ยังช่วยสนับสนุนในการตัดสินใจสำหรับการกำหนดค่าพารามิเตอร์บางส่วนของเคเนียร์เรสเนเบอร์

5.3 ปัญหาและข้อเสนอแนะ

งานวิจัยเกิดขึ้นเนื่องจากปัญหาการใช้งานเคเนียร์เรสเนเบอร์ ที่จำเป็นต้องกำหนดค่า เค และมาตรวัดระยะทางก่อนการจำแนกข้อมูล ซึ่งในงานวิจัยนี้เลือกใช้การทดสอบแบบมาร์เดียในการในการพิจารณา ซึ่งในอนาคตอาจมีการใช้ค่าอื่น ๆ ของมาร์เดียในการร่วมพิจารณาเพื่อกำหนดค่า

รายการอ้างอิง

นลินี โสพัศสถิตย. (2555). **การใช้ระบบแนะนำสนับสนุนการตัดสินใจ**. [ออนไลน์]. ได้จาก : http://www.ssruir.ssru.ac.th/bitstream/ssruir/659/1/083-55.pdf

Bhattacharya, G., Ghosh, K., and Chowdhury, A.S. (2014). Test Point Specific k Estimation for kNN Classifier. **In Proceedings of the 22nd International Conference on Pattern Recognition** (pp 1478-1483).

Bunheang, T., Jung K.H., and Sejong, O. (2014). A Machine Learning Approach for Specification of Spinal Cord Injuries Using Fractional Anisotropy Values Obtained from Diffusion Tensor Images. **Computational and Mathematical Methods in Medicine.** 2014: Article ID 276589.

Cantrell, C. D. (2000). Modern Mathematical Methods for Physicists and Engineers. **Cambridge University Press**.

Deza, E., Deza, M. M. (2009). Encyclopedia of Distances. **Springer**. Berlin. : 94.

Ghosh, A. K. (2006). On Optimum Choice of *k* in Nearest Neighbor Classification. **Computational Statistics & Data Analysis.** 50: 3113-3123.

Hand, D. J. and Vinciotti, V. (2003). Choosing k for Two-Class Nearest Neighbor Classifiers with Unbalanced Classes. **Pattern Recognition**. 24: 1555-1562.

Hu, S-b. and Shao, P. (2012). Improved Nearest Neighbor Interpolators Based on Confidence Region in Medical Image Registration. **Biomedical Signal Processing and Control.** 7: 525-536.

Hulett, C., Hall, A., and Qu G. (2012) Dynamic Selection of k Nearest Neighbors in Instance-Based Learning**.** Information Reuse and Integration. **:** 85-92.

Krause, E. F. (1987). Taxicab Geometry. **Dover**.

Lee, T. and Ouarda, T. B.M.J. (2011). Identification of Model Order and Number of Neighbors for k-Nearest Neighbor Resampling. **Journal of Hydrology.** 404: 136-145.

Mardia, K. V. (1970). Measures of Multivariate Skewness and Kurtosis with Applications. **Biometrika**. 57(3): 519-530.

Mardia, K. V. (1974). Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies. **Sankhya Ser.** 36(2): 115-128.

Medjahed, S. A., Saadi, T. A., and Benyettou (2013). A Breast Cancer Diagnosis by Using k-Nearest Neighbor with Different Distances and Classification Rules. **International Journal of Computer Applications**. 62(1): 1-5.

Peterson, L. E. (2009). k-Nearest Neighbor. **Scholarpedia**, 4(2): 1883.

Tan, P.-N., Steinbach, M., and V. Kumar (2005). Introduction to Data Mining. **Bulletin of the IEEE Computer Society Technical Committee on Data Engineering**. Addison-Wesley, Chapter 8: 500.

Yoon, J.W. and Friel, N. (2015). Efficient Model Selection for Probabilistic K Nearest Neighbour Classification. **Neurocomputing.** 149: 1098–1108.

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภาคผนวก ก

การใช้งานโปรแกรม

การใช้งานโปรแกรม

เนื้อหาในส่วนนี้เป็นการอธิบายเกี่ยวกับขั้นตอนวิธีการต่าง ๆ ของการทำการทดลอง ไม่ว่าจะเป็นการเตรียมข้อมูลในการทดลอง การใช้รหัสต้นของโปรแกรม การใช้งานเครื่องมือ ซึ่งสามารถดำเนินการได้ดังต่อไปนี้

1. การเตรียมข้อมูล

การเตรียมข้อมูลเป็นการจัดการกับข้อมูล เพื่อให้ข้อมูลนั้น ๆ พร้อมที่จะใช้งาน เพราะหากไม่มีการเตรียมข้อมูลเมื่อนำข้อมูลไปใช้งานอาจทำให้เกิดข้อผิดพลาดจากตัวข้อมูลเองไม่ว่าจะเป็นการที่ข้อมูลมีบางช่วงบางตอนขาดหายไป หรือมีอักษรพิเศษที่ไม่สามารถประมวลผลได้ การเตรียมข้อมูลเพื่อใช้งานจึงเป็นส่วนที่จำเป็น ซึ่งในงานวิจัยนี้ได้ใช้เคเนียร์เรสเนเบอร์ในการจำแนก โดยที่การจำแนกข้อมูลบางคอลัมน์ไม่สามารถทำการวัดระยะทางได้ เช่น คอลัมน์เพศ ดังรูปที่ ก.1

ข้อมูลดั้งเดิม

Sex	Fbs
1	0
0	0
1	0
1	0
0	0
1	0
1	1
1	0
1	0

→

Dummy Variable

Sex=F	Sex=M	Fbs=T	Fbs=F
1	0	0	1
0	1	0	1
1	0	0	1
1	0	0	1
0	1	0	1
1	0	0	1
1	0	1	0
1	0	0	1
1	0	0	1

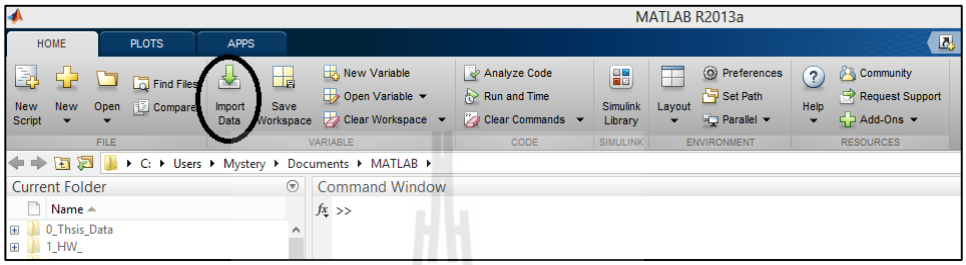
รูปที่ ก.1 การเตรียมข้อมูล

เมื่อต้องการทำให้คอลัมน์ที่ไม่สามารถวัดระยะทางได้สามารถวัดได้ จึงมีการแปลงข้อมูลให้อยู่ในรูป Dummy Variable ซึ่งเป็นการเพิ่มคอลัมน์ให้แก่ข้อมูลตามจำนวนค่าที่เป็นไปได้ เช่น เพศ เดิมมี 2 เพศ คือ เพศชาย (M) และ เพศหญิง (F) จากรูปที่ ก.1 ในข้อมูลดั้งเดิมแทนเพศหญิงด้วย 1 และ เพศชายด้วย 0 เมื่อทำการแปลงจำนวนคอลัมน์จะเพิ่มขึ้นเป็น 2 คอมลัมน์จากเดิมที่มีเพียงคอลัมน์ Sex เพียงอย่างเดียว

1.1 การนำข้อมูลเข้าโปรแกรม MATLAB

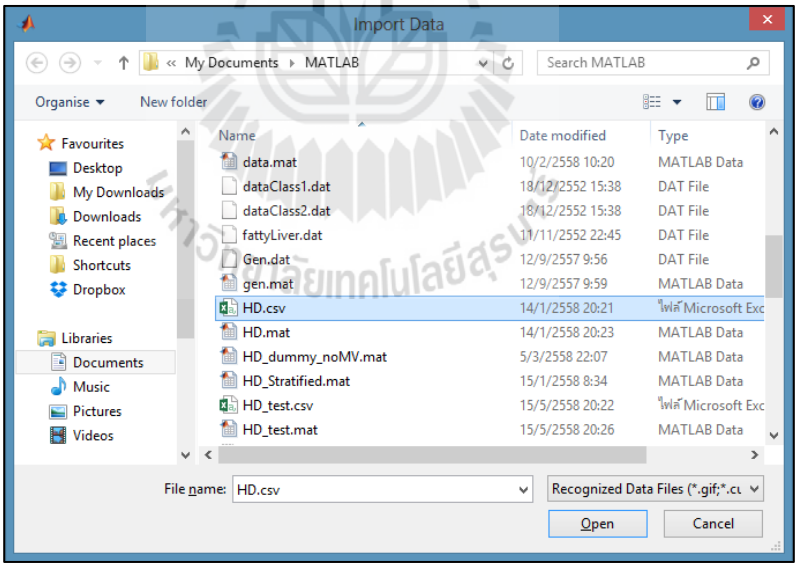
เนื่องจากงานวิจัยนี้ใช้เครื่องมือ MATLAB เพื่อใช้ในการจำแนกโดยเคเนียร์เรสเนเบอร์ การเตรียมข้อมูลสำหรับ MATLAB นั้นสามารถทำได้ดังขั้นตอนต่อไปนี้

1. เริ่มการทำงานของโปรแกรม MATLAB
2. ทำการ Import Data ดังรูป ก.2



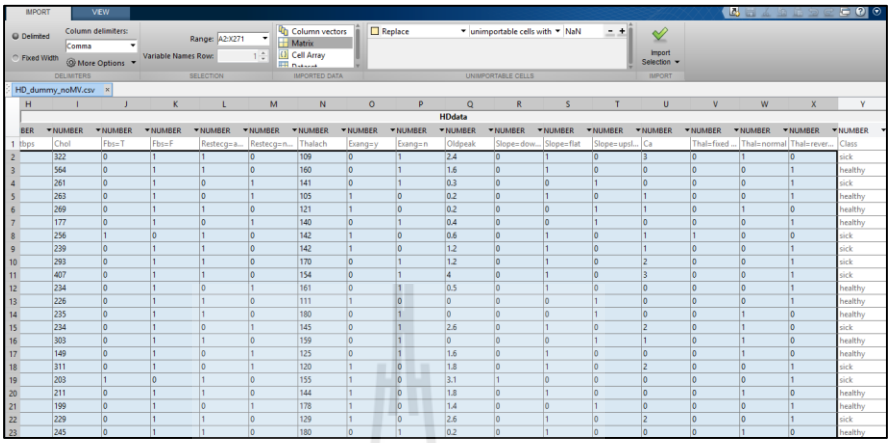
รูปที่ ก.2 การ Import Data

3. เลือกไฟล์ข้อมูลที่ต้องการ ดังรูปที่ ก.3



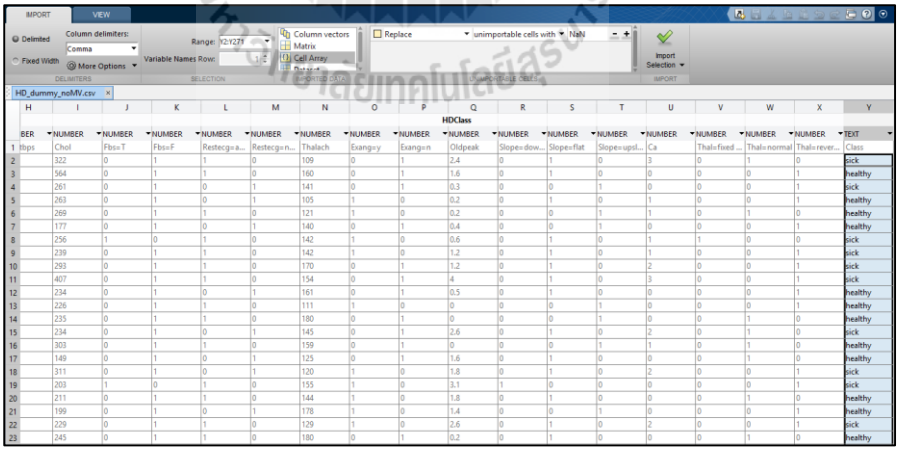
รูปที่ ก.3 เลือกไฟล์ข้อมูล

4. เมื่อทำการเลือกไฟล์ข้อมูลที่ต้องการแล้ว จะมีหน้าต่างข้อมูลปรากฏขึ้น ซึ่งเป็นหน้าต่างสำหรับการแยกระหว่างข้อมูลกับคลาสเป้าหมาย ดังรูปที่ ก.4 และ ก.5



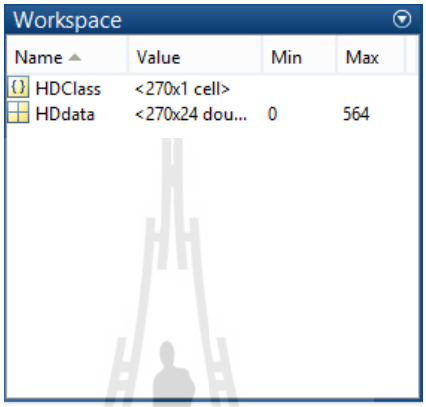
รูปที่ ก.4 เลือกส่วนของข้อมูลสำหรับการ Import

จากรูปที่ ก.4 จะเป็นการเลือกส่วนของข้อมูลโดยส่วนของข้อมูลจะมีลักษณะการ Imported Data เป็น Matrix ซึ่งต่างกับคลาสเป้าหมายที่จะมีลักษณะ Imported Data เป็น Cell Array ดังรูปที่ ก.5



รูปที่ ก.5 เลือกส่วนของคลาสเป้าหมายสำหรับการ Import

หลังจากกำหนดข้อมูลแล้วว่าเป็นส่วนของข้อมูลกับส่วนของคลาสเป้าหมายสามารถ กดที่เครื่องหมายถูกมุมบนขวาเพื่อทำการ Import Data เข้าสู่ MATLAB ได้ทันที เมื่อข้อมูลถูก Import เข้าสู่ MATLAB ข้อมูลจะถูกแสดงรายละเอียดที่ Workspace ของ MATLAB ซึ่งจะมีลักษณะดังรูปที่ ก.6



รูปที่ ก.6 Workspace ของ MATLAB

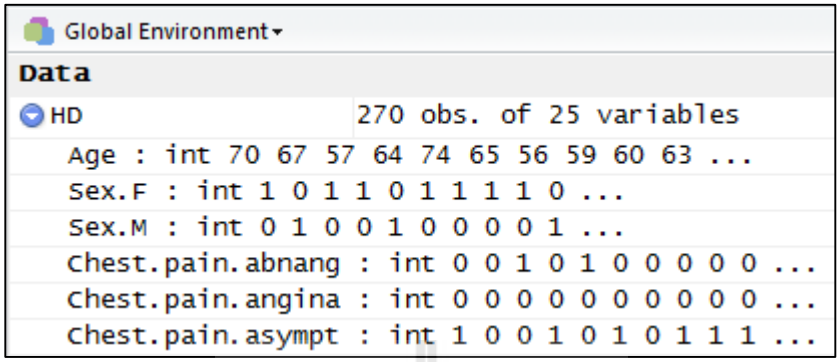
สำหรับการเตรียมข้อมูลเข้าสู่โปรแกรม MATLAB เมื่อส่วนของข้อมูลและคลาสเป้าหมายมีลักษณะดัง รูปที่ ก.6 ก็ถือว่านำเข้าข้อมูลได้สำเร็จ หลังจากนี้สามารถดำเนินการกับข้อมูลต่อได้ทันที

1.2 การนำข้อมูลเข้าโปรแกรม RStudio

เนื่องจากงานวิจัยนี้ใช้โปรแกรม RStudio ในการทดสอบมาร์เดีย ก่อนที่จะทำการทดสอบได้จำเป็นต้องมีการนำข้อมูลเข้าสู่โปรแกรม Rstudio ก่อน ซึ่งสามารถดำเนินการได้ดังนี้

1. เริ่มโปรแกรม RStudio
2. ใช้คำสั่ง read.csv ในการนำข้อมูลเข้า โดยที่พารามิเตอร์ของเป็นดังต่อไปนี้ read.csv(“Path/ชื่อไฟล์.csv”) ตัวอย่างเช่น HD <- read.csv(“C:/Data/HD.csv”) ซึ่งหมายถึงเก็บข้อมูลไว้ในตัวแปร

เมื่อนำข้อมูลเข้าได้สำเร็จ รายละเอียดต่าง ๆ ของข้อมูลจะอยู่ในส่วน Global Environment ดังรูปที่ ก.7



รูปที่ ก.7 ลักษณะข้อมูลเมื่อนำเข้าสู่ RStudio

2. การจำแนกข้อมูลโดยเคเนียร์เรสเนเบอร์ใน MATLAB

เนื่องจากงานวิจัยนี้มีการใช้งานเคเนียร์เรสเนเบอร์ใน MATLAB จึงจัดทำการอธิบายการทำงานบางส่วนของการใช้งานเคเนียร์เรสเนเบอร์ใน MATLAB ขึ้น โดยสามารถดำเนินการได้ดังนี้ หลังจากการนำข้อมูลเข้าสู่ MATLAB ดังรูปที่ ก.6 ซึ่งแบ่งออกเป็น 2 ส่วนคือส่วนของข้อมูลและส่วนของคลาสเป้าหมาย การใช้งานเคเนียร์เรสเนเบอร์ใน MATLAB นั้นสามารถใช้งานได้โดยฟังก์ชัน ClassificationKNN.fit(HDdata, HDClass, 'Distance', 'euclidean', 'NumNeighbors', 1) ซึ่งพารามิเตอร์ในส่วนต่าง ๆ มีความหมายดังต่อไปนี้

HDdata คือส่วนของข้อมูล

HDClass คือส่วนของคลาสเป้าหมาย

Distance คือมาตรวัดระยะทาง

'euclidean' คือกำหนดมาตรวัดระยะทางเป็น Euclidean

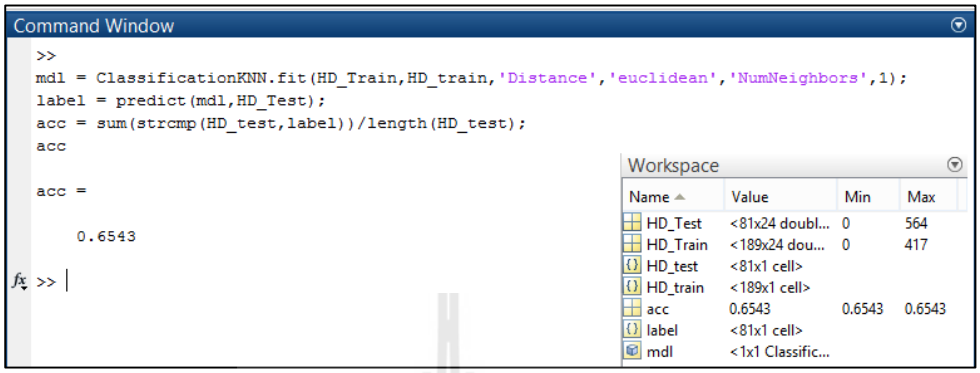
NumNeighbors คือจำนวน เค ของเคเนียร์เรสเนเบอร์

1 คือกำหนดให้ค่า เค ของเคเนียร์เรสเนเบอร์เป็น 1

เมื่อมีข้อมูลทดสอบที่ต้องการทดสอบกับเคเนียร์เรสเนเบอร์สามารถทำได้โดยใช้ฟังก์ชัน predict(mdl,HDdatats) ซึ่ง mdl คือส่วนของ ClassificationKNN.fit และ HDdatats คือส่วนของข้อมูลทดสอบ โดยฟังก์ชัน predict จะเป็นการทำนายคลาสแก่ข้อมูลทดสอบ

เมื่อทำการวัดประสิทธิภาพด้วยค่าความแม่นสามารถทำได้โดยใช้ผลรวมของการเปรียบเทียบคลาสเป้าหมายของข้อมูลทดสอบกับการทำนาย และหารด้วยจำนวนข้อมูลทดสอบ

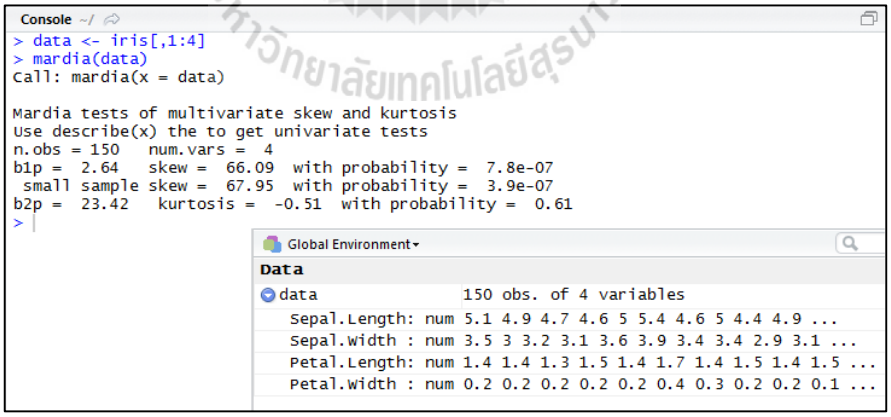
ทั้งหมด ดังการใช้ฟังก์ชันต่อไปนี้ sum(strcmp(tsC,label))/length(tsC) โดยกระบวนการข้างต้นเป็นดังรูปที่ ก.8



รูปที่ ก.8 ตัวอย่างการใช้งานเคเนียร์เรสเนเบอร์

3. การทดสอบมาร์เดียใน RStudio

การเรียกใช้งานมาร์เดียใน RStudio สามารถทำได้โดยการเพิ่ม Library ที่เกี่ยวข้องลงไป โดยในที่นี้เมื่อต้องการทดสอบมาร์เดียสามารถเรียกใช้ Library ที่มีชื่อว่า psych หลังจากนั้นจะสามารถใช้งานฟังก์ชัน mardia(data) ได้ ซึ่ง data คือข้อมูลที่ต้องการทดสอบ ดังตัวอย่างการใช้งานอย่างง่ายดังรูปที่ ก.9



รูปที่ ก.9 ตัวอย่างการใช้งานการทดสอบมาร์เดียด้วย RStudio

ภาคผนวก ข

รหัสต้นฉบับของโปรแกรม

มหาวิทยาลัยเทคโนโลยีสุรนารี

โปรแกรมวัดประสิทธิภาพการทำงาน เค เนียร์เรสเนเบอร์ด้วยมาตรวัดระยะทางต่าง ๆ

```
load('xDatasets.mat')
BC = CalKnn(BC_Train,BC_train,BC_Test,BC_test)
PM = CalKnn(PM_Train,PM_train,PM_Test,PM_test)
HD = CalKnn(HD_Train,HD_train,HD_Test,HD_test)
AM = CalKnn(AM_Train,AM_train,AM_Test,AM_test)
TR = CalKnn(TR_Train,TR_train,TR_Test,TR_test)

function [ result ] = CalKnn( trD, trC, tsD, tsC )
D = {'euclidean','cityblock','cosine','correlation'};
%,'mahalanobis','cosine','hamming','jaccard',chebychev
nDis = size(D,2);
result = cell(nDis,11);
for idist=1:nDis
   a= [1];
   a=[a, round((1:10)*(size(tsD,1)/10))];
   kInd = 2;
   result{idist,1} = D{idist};
   for ice=a
      mdl = ClassificationKNN.fit(trD,trC,'Distance',D{idist},'NumNeighbors',ice);
      label = predict(mdl,tsD);
      acc = sum(strcmp(tsC,label))/length(tsC);
      result{idist,kInd} = acc;
      kInd = kInd +1;
   end
end
end
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

โปรแกรมการทดสอบแบบมาร์เดียกับข้อมูลทางการแพทย์

```
### Library/Sampling Function ###
library("psych", lib.loc="~/R/win-library/3.2")
library("sampling", lib.loc="~/R/win-library/3.2")

preprocess <- function(xdata){
        tmp <- vector()
        for(i in 1:dim(xdata)[2]){
                if(length(unique(xdata[,i]))>1) {
                        tmp <- c(tmp,i)
                }
        }
        return(tmp)
}

set.seed(0)
stsampling <- function(X,target,test){
output <- list()
idtrain <- strata(X,stratanames=target,size=table(X[,target])*(1-test), method="srswor")$ID_unit
output$train <- X[idtrain,]
output$test <- X[(idtrain*-1),]
return (output)
}

### HD ###
heart_r <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/HD.csv")
heart <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-
58_Paper/HD_dummy_noMV.csv")
HD <- stsampling(heart,"Class",0.3)
```

```
set.seed(5)
HD_id20 <- strata(HD$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tHD <- heart_r[row.names(HD$test),]
tHD$Class <- NULL
mardia(tHD[HD_id20,])
set.seed(6)
HD_id20 <- strata(HD$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tHD <- heart_r[row.names(HD$test),]
tHD$Class <- NULL
mardia(tHD[HD_id20,])
set.seed(13)
HD_id20 <- strata(HD$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tHD <- heart_r[row.names(HD$test),]
tHD$Class <- NULL
mardia(tHD[HD_id20,])

### BC ###
breast2 <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/breast-
cancer-wisconsin.csv", na.string = "?")
BC <- stsampling(breast2,"X2.1",0.3)

set.seed(5)
BC_id20 <- strata(BC$test,stratanames="X2.1",size=c(10,10), method="srswor")$ID_unit
tBC <- breast2[row.names(BC$test),]
tBC$X2.1 <- NULL
mardia(tBC[BC_id20,])
set.seed(6)
BC_id20 <- strata(BC$test,stratanames="X2.1",size=c(10,10), method="srswor")$ID_unit
tBC <- breast2[row.names(BC$test),]
tBC$X2.1 <- NULL
```

```
mardia(tBC[BC_id20,])
set.seed(13)
BC_id20 <- strata(BC$test,stratanames="X2.1",size=c(10,10), method="srswor")$ID_unit
tBC <- breast2[row.names(BC$test),]
tBC$X2.1 <- NULL
mardia(tBC[BC_id20,])

### PM ###
pima <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/pima-
indians-diabetes.txt")
PM <- stsampling(pima,"X1",0.3)

set.seed(5)
PM_id20 <- strata(PM$test,stratanames="X1",size=c(10,10), method="srswor")$ID_unit
tPM <- pima[row.names(PM$test),]
tPM$X2.1 <- NULL
mardia(tPM[PM_id20,])
set.seed(6)
PM_id20 <- strata(PM$test,stratanames="X1",size=c(10,10), method="srswor")$ID_unit
tPM <- pima[row.names(PM$test),]
tPM$X2.1 <- NULL
mardia(tPM[PM_id20,])
set.seed(13)
PM_id20 <- strata(PM$test,stratanames="X1",size=c(10,10), method="srswor")$ID_unit
tPM <- pima[row.names(PM$test),]
tPM$X2.1 <- NULL
mardia(tPM[PM_id20,])

### TR ###
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
thyroid2c <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/ann-thyroid.full2c.csv")
TR <- stsampling(thyroid2c,"Class",0.3)

set.seed(5)
TR_id20 <- strata(TR$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tTR <- thyroid2c[row.names(TR$test),]
tTR$Class <- NULL
ppTR <- preprocess(tTR[TR_id20,])
mardia(tTR[TR_id20,ppTR])
set.seed(6)
TR_id20 <- strata(TR$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tTR <- thyroid2c[row.names(TR$test),]
tTR$Class <- NULL
ppTR <- preprocess(tTR[TR_id20,])
mardia(tTR[TR_id20,ppTR])
set.seed(13)
TR_id20 <- strata(TR$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tTR <- thyroid2c[row.names(TR$test),]
tTR$Class <- NULL
ppTR <- preprocess(tTR[TR_id20,])
mardia(tTR[TR_id20,ppTR])

### AM ###5#13#16
asthma_r <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/PA_noMV2c.csv")
asthma <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/PA_dummy_noMV2c.csv")
asthma_r[,"PA_level"] <- relevel(asthma_r[,"PA_level"],"low-mod")
asthma[,"PA_level"] <- relevel(asthma[,"PA_level"],"low-mod")
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
AM <- stsampling(asthma_r,"PA_level",0.3)


set.seed(1)
AM_id20 <- strata(AM$test,stratanames="PA_level",size=c(10,10), method="srswor")$ID_unit
tAM <- asthma_r[row.names(AM$test),]
tAM$PA_level <- NULL
ppAM <- preprocess(tAM[AM_id20,])
mardia(tAM[AM_id20,ppAM])
set.seed(2)
AM_id20 <- strata(AM$test,stratanames="PA_level",size=c(10,10), method="srswor")$ID_unit
tAM <- asthma_r[row.names(AM$test),]
tAM$PA_level <- NULL
ppAM <- preprocess(tAM[AM_id20,])
mardia(tAM[AM_id20,ppAM])
set.seed(3)
AM_id20 <- strata(AM$test,stratanames="PA_level",size=c(10,10), method="srswor")$ID_unit
tAM <- asthma_r[row.names(AM$test),]
tAM$PA_level <- NULL
ppAM <- preprocess(tAM[AM_id20,])
mardia(tAM[AM_id20,ppAM])
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

โปรแกรมวัดประสิทธิภาพการทำงาน เค เนียร์เรสเนเบอร์ด้วยมาตรวัดระยะทางต่าง ๆ

```
load('xDatasets.mat')
BC = CalKnn(BC_Train,BC_train,BC_Test,BC_test)
PM = CalKnn(PM_Train,PM_train,PM_Test,PM_test)
HD = CalKnn(HD_Train,HD_train,HD_Test,HD_test)
AM = CalKnn(AM_Train,AM_train,AM_Test,AM_test)
TR = CalKnn(TR_Train,TR_train,TR_Test,TR_test)

function [ result ] = CalKnn( trD, trC, tsD, tsC )
D = {'euclidean','cityblock','cosine','correlation'};
%,'mahalanobis','cosine','hamming','jaccard',chebychev
nDis = size(D,2);
result = cell(nDis,11);
for idist=1:nDis
   a= [1];
   a=[a, round((1:10)*(size(tsD,1)/10))];
   kInd = 2;
   result{idist,1} = D{idist};
   for ice=a
      mdl = ClassificationKNN.fit(trD,trC,'Distance',D{idist},'NumNeighbors',ice);
      label = predict(mdl,tsD);
      acc = sum(strcmp(tsC,label))/length(tsC);
      result{idist,kInd} = acc;
      kInd = kInd +1;
   end
end
end
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

โปรแกรมการทดสอบแบบมาร์เดียกับข้อมูลทางการแพทย์

```
### Library/Sampling Function ###
library("psych", lib.loc="~/R/win-library/3.2")
library("sampling", lib.loc="~/R/win-library/3.2")

preprocess <- function(xdata){
        tmp <- vector()
        for(i in 1:dim(xdata)[2]){
                if(length(unique(xdata[,i]))>1) {
                        tmp <- c(tmp,i)
                }
        }
        return(tmp)
}

set.seed(0)
stsampling <- function(X,target,test){
output <- list()
idtrain <- strata(X,stratanames=target,size=table(X[,target])*(1-test), method="srswor")$ID_unit
output$train <- X[idtrain,]
output$test <- X[(idtrain*-1),]
return (output)
}

### HD ###
heart_r <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/HD.csv")
heart <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-
58_Paper/HD_dummy_noMV.csv")
HD <- stsampling(heart,"Class",0.3)
```

```
set.seed(5)
HD_id20 <- strata(HD$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tHD <- heart_r[row.names(HD$test),]
tHD$Class <- NULL
mardia(tHD[HD_id20,])
set.seed(6)
HD_id20 <- strata(HD$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tHD <- heart_r[row.names(HD$test),]
tHD$Class <- NULL
mardia(tHD[HD_id20,])
set.seed(13)
HD_id20 <- strata(HD$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tHD <- heart_r[row.names(HD$test),]
tHD$Class <- NULL
mardia(tHD[HD_id20,])

### BC ###
breast2 <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/breast-
cancer-wisconsin.csv", na.string = "?")
BC <- stsampling(breast2,"X2.1",0.3)

set.seed(5)
BC_id20 <- strata(BC$test,stratanames="X2.1",size=c(10,10), method="srswor")$ID_unit
tBC <- breast2[row.names(BC$test),]
tBC$X2.1 <- NULL
mardia(tBC[BC_id20,])
set.seed(6)
BC_id20 <- strata(BC$test,stratanames="X2.1",size=c(10,10), method="srswor")$ID_unit
tBC <- breast2[row.names(BC$test),]
tBC$X2.1 <- NULL
```

```
mardia(tBC[BC_id20,])
set.seed(13)
BC_id20 <- strata(BC$test,stratanames="X2.1",size=c(10,10), method="srswor")$ID_unit
tBC <- breast2[row.names(BC$test),]
tBC$X2.1 <- NULL
mardia(tBC[BC_id20,])

### PM ###
pima <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/pima-
indians-diabetes.txt")
PM <- stsampling(pima,"X1",0.3)

set.seed(5)
PM_id20 <- strata(PM$test,stratanames="X1",size=c(10,10), method="srswor")$ID_unit
tPM <- pima[row.names(PM$test),]
tPM$X2.1 <- NULL
mardia(tPM[PM_id20,])
set.seed(6)
PM_id20 <- strata(PM$test,stratanames="X1",size=c(10,10), method="srswor")$ID_unit
tPM <- pima[row.names(PM$test),]
tPM$X2.1 <- NULL
mardia(tPM[PM_id20,])
set.seed(13)
PM_id20 <- strata(PM$test,stratanames="X1",size=c(10,10), method="srswor")$ID_unit
tPM <- pima[row.names(PM$test),]
tPM$X2.1 <- NULL
mardia(tPM[PM_id20,])

### TR ###
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
thyroid2c <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/ann-thyroid.full2c.csv")
TR <- stsampling(thyroid2c,"Class",0.3)

set.seed(5)
TR_id20 <- strata(TR$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tTR <- thyroid2c[row.names(TR$test),]
tTR$Class <- NULL
ppTR <- preprocess(tTR[TR_id20,])
mardia(tTR[TR_id20,ppTR])
set.seed(6)
TR_id20 <- strata(TR$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tTR <- thyroid2c[row.names(TR$test),]
tTR$Class <- NULL
ppTR <- preprocess(tTR[TR_id20,])
mardia(tTR[TR_id20,ppTR])
set.seed(13)
TR_id20 <- strata(TR$test,stratanames="Class",size=c(10,10), method="srswor")$ID_unit
tTR <- thyroid2c[row.names(TR$test),]
tTR$Class <- NULL
ppTR <- preprocess(tTR[TR_id20,])
mardia(tTR[TR_id20,ppTR])

### AM ###5#13#16
asthma_r <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/PA_noMV2c.csv")
asthma <- read.csv("E:/Dropbox/Public/วิชาการ/58-1/Advance Datamining/1-58_Paper/PA_dummy_noMV2c.csv")
asthma_r[,"PA_level"] <- relevel(asthma_r[,"PA_level"],"low-mod")
asthma[,"PA_level"] <- relevel(asthma[,"PA_level"],"low-mod")
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
AM <- stsampling(asthma_r,"PA_level",0.3)


set.seed(1)
AM_id20 <- strata(AM$test,stratanames="PA_level",size=c(10,10), method="srswor")$ID_unit
tAM <- asthma_r[row.names(AM$test),]
tAM$PA_level <- NULL
ppAM <- preprocess(tAM[AM_id20,])
mardia(tAM[AM_id20,ppAM])
set.seed(2)
AM_id20 <- strata(AM$test,stratanames="PA_level",size=c(10,10), method="srswor")$ID_unit
tAM <- asthma_r[row.names(AM$test),]
tAM$PA_level <- NULL
ppAM <- preprocess(tAM[AM_id20,])
mardia(tAM[AM_id20,ppAM])
set.seed(3)
AM_id20 <- strata(AM$test,stratanames="PA_level",size=c(10,10), method="srswor")$ID_unit
tAM <- asthma_r[row.names(AM$test),]
tAM$PA_level <- NULL
ppAM <- preprocess(tAM[AM_id20,])
mardia(tAM[AM_id20,ppAM])
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภาคผนวก ค

บทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

มหาวิทยาลัยเทคโนโลยีสุรนารี

รายชื่อบทความวิจัยที่ได้รับการตีพิมพ์เผยแพร่ในระหว่างศึกษา

P. Teerarassamee, K. Kerdprasop, N. Kerdprasop. (2015). **The Methodology to Find Appropriate k for k-Nearest Neighbor Classification with Medical Datasets**. In Proceedings of the 9th SOUTH EAST ASIAN TECHNICAL UNIVERSITY CONSORTIUM (SEATUC) SYMPOSIUM. Nakhon Ratchasima, Thailand. 27 - 30 July 2015

K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, N. Kerdprasop (2015). **An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm.** Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015), Kitakyushu, Japan, 28-31 March, pp.280-285.

มหาวิทยาลัยเทคโนโลยีสุรนารี

THE METHODOLOGY TO FIND APPROPRIATE K FOR K-NEAREST NEIGHBOR CLASSIFICATION WITH MEDICAL DATASETS

Pongsakorn Teerarassamee, Kittisak Kerdprasop, Nittaya Kerdprasop
School of Computer Engineering, Suranaree University of Technology, Thailand

ABSTRACT

This research studies the problem of distance-based data classification using k-nearest neighbor algorithm for classifying medical datasets. The classification of data needs to define value k of the most k nearest data points and then classifying new data point to be the same class as the majority value among the k data. If the choice of value k is not suitable, then the accuracy of the classification is lower than it should be. On the contrary, if, user is defines a lot of value of k, it will result in a very slow process. This paper presents the guideline regarding how to select an appropriate value k to the medical field, by considering the nature of the classes and instances. A different distance metric also results in different level of accuracy. In this study, we perform k-nearest neighbor classification using 8 different distance measurement. The suitable distance metric and appropriate k value and reported as a guideline to the user.

1. INTRODUCTION

The k-nearest neighbor algorithm (kNN) is a method of classification. It is categorized as supervised learning. That means the data already know the answer of Data Classification is the process of learning from data is the answer. Techniques such as the widely used, whether it is medical (Hu & Shao, 2012) or the Hydrology and Meteorology (Lee & Ouarda, 2011). The configuration value k is the scope or in the data analysis. By setting value k are the results on the use of resources to analyze data. The time it takes to process if the value of k, high processing will take longer. This research was presented on how to find the appropriate value k is applied to medical data. Such as data on asthma patients at different levels. Data for heart patients and breast cancer, etc. There is also speculation about the distance. (Medjahed, Saadi, & Benyettou, 2013). Different again In addition to the appropriate value k to reduce the processing time and also added the validity as well. In this study, Selected to 8 different distances because of the nature of distance often give an not equal, so that is why this study, those selected distance.

The people who want to use the k-Nearest Neighbor classification methods from this research can be used to configure value k of k-Nearest Neighbor immediately.

2. ANALYSIS

In this studied, purpose is to determine value k appropriate, taking into consideration the type of distance because the distance is often different to the validity different. And characteristics that affect the configuration information as well. We using 8 different distances for compare performance.

- Euclidean Distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (1)$$

Let $d(p,q)$ is distance of Euclidean. p_i and q_i is points in dimension.

Euclidean Distance (Deza & Deza, 2009). is popular distance because easy to understand. By measuring the distance of the two points.

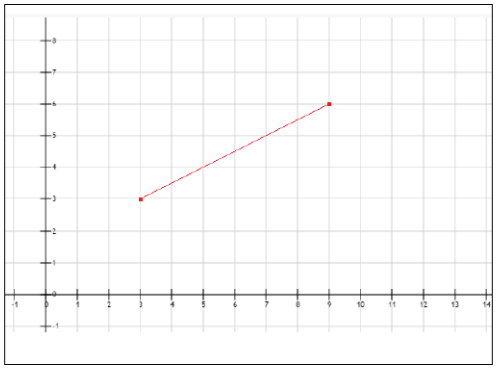


Fig. 4 Euclidean Distance.

- Cityblock (or Manhattan) Distance

$$d(p,q) = \sum_{i=1}^{n} |p_i - q_i| \quad (2)$$

Let $d(p,q)$ is distance of Cityblock. (Krause, 1987) p_i and q_i is points in dimension.

Cityblock Distance is driving route distance.

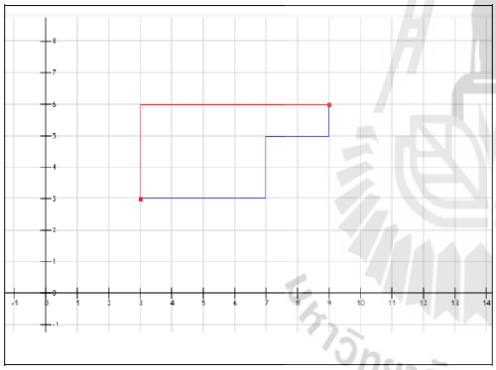


Fig. 5 Cityblock (or Manhattan) Distance (red or blue line).

- Chebyshev Distance

$$d(p,q) = \max_i(|p_i - q_i|) \quad (3)$$

Let $d(p,q)$ is distance of Chebyshev. (Cantrell, 2000) p_i and q_i is points in dimension.

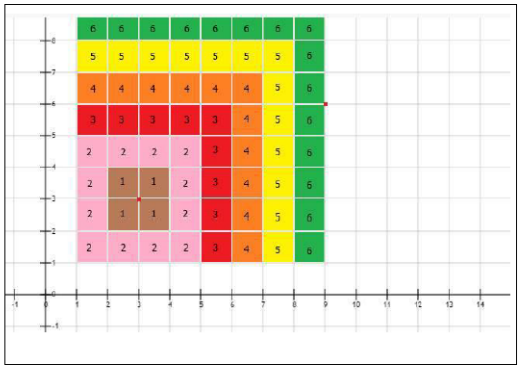


Fig. 6 Chebyshev Distance.

- Correlation Distance

$$d(p,q) = 1 - \frac{cov(p,q)}{std(p)*std(q)} \quad (4)$$

$$cov(p,q) = \sum_{j=1}^{k}(p_j - \bar{p}) * (q_j - \bar{q}) \quad (5)$$

$$std(p) = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(p_j - \bar{p})^2} \quad (6)$$

$$\bar{p} = \frac{1}{k}\sum_{j=1}^{k} p_j \quad (7)$$

Let $d(p,q)$ is distance of Correlation. (Pearson, 1895) p_j and q_j is points in dimension. $cov(p,q)$ is covariance. and std(p) is standard deviation.

- Cosine Distance

$$d(p,q) = 1 - \frac{\sum_{i=1}^{n} p_i * q_i}{\sqrt{\sum_{i=1}^{n}(p_i)^2} * \sqrt{\sum_{i=1}^{n}(q_i)^2}} \quad (8)$$

Let $d(p,q)$ is distance of Cosine. (Singhal, 2001) p_i and q_i is points in dimension.

- Hamming Distance

$$d(p,q) = \sum_{k=0}^{n-1}[y_{p,k} \neq y_{q,k}] \quad (9)$$

Let $d(p,q)$ is distance of Hamming. (Hamming, 1950) $y_{p,k}$ and $y_{q,k}$ is items or points in dimension.

- Jaccard Distance

$$d(p,q) = 1 - \frac{J_{in}}{J_p + J_q + J_{in}} \quad (10)$$

Let $d(p,q)$ is distance of Jaccard. (Jaccard, 1901) J_{in} is number of intersection between item p and item q. J_p is number of item p and J_q is number of item q.

มหาวิทยาลัยเทคโนโลยีสุรนารี

3. EXPERIMENT

In experiments, we used 2 data sets from the UCI Machine Learning Repository and 1 data set from the Maharat Nakhon Ratchasima Hospital on 8 Nov 2014 (Table 1).

We compare the 8 distances to find the appropriate k value and the best distance for k-Nearest Neighbor. The classification performances of all cases were measured by ten-fold cross validation.

Table 1 Basic information of data sets.

Name	Instances	Attributes	Classes
Breast Cancer Wisconsin	683	10	2
Heart Disease	270	13	2
Asthma	698	13	3

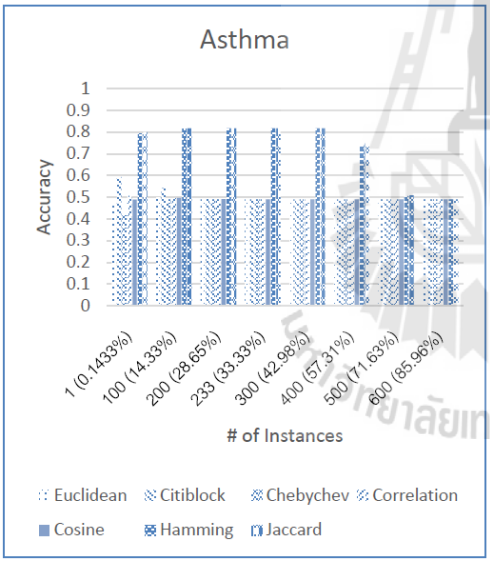


Fig. 1 Classification asthma with 7 different distances.

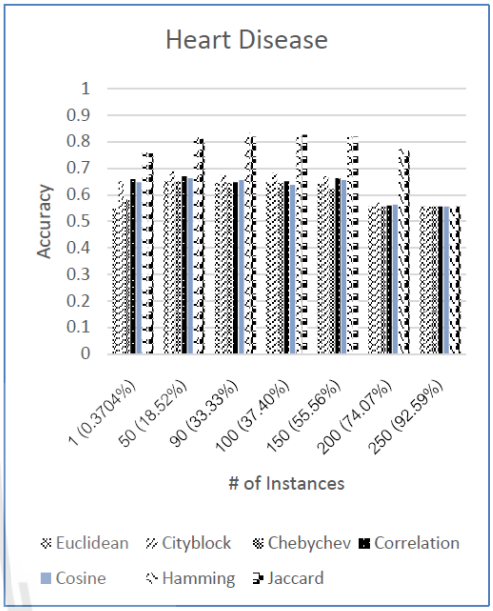


Fig. 2 Classification heart disease with 7 different distances.

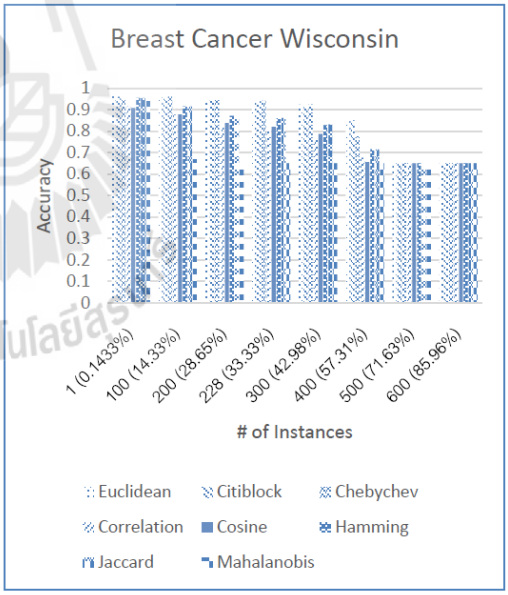


Fig. 3 Classification breast cancer wisconsin with 8 different distances.

The results showed in Fig.1 – 3 is a relationship between accuracy and number of instances. From asthma, It can be seen that Hamming Distance and Jaccard Distance have accuracy high value and is similar to heart disease. In breast cancer wisconsin when k = 1 will give the accuracy in all the instances.

In asthma datasets and heart disease datasets can't use Mahalanobis Distance because the calculation requires covariance symmetry but all above datasets have covariance asymmetric.

4. CONCLUSION

In this study, Using 8 distances with k-Nearest Neighbor and each will have a different distance away from the results obtained are found to classes and the number of instances affect the configuration k with the following characteristics.

If there are 3 classes and a number of instances value of k that will be in the range of 30% of instances with odometer Hamming Distance and Jaccard Distance.

If there are 2 classes of data and the number of instances value of k that will be in the range of 20% of instances with all above distance but except the Hamming Distance and Jaccard Distance.

If the data is Instances least 2 classes and the appropriate k can be used as k=1 immediately.

REFERENCES

Medjahed, S.A., Saadi, T.A., and Benyettou., 2013. A Breast Cancer Diagnosis by Using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Applications. 62(1): 1-5.

Hu, S-b. and Shao, P., 2012. Improved Nearest Neighbor Interpolators Based on Confidence Region in Medical Image Registration. Biomedical Signal Processing and Control. 7: 525-536.

Lee, T. and Ouarda, T. B.M.J., 2011. Identification of Model Order and Number of Neighbors for k-Nearest Neighbor Resampling. Journal of Hydrology. 404: 136-145.

Deza, E., Deza, M. M., 2009. Encyclopedia of Distances. Springer. : 94.

Krause, E. F., 1987. Taxicab Geometry. Dover.

Cantrell, C. D., 2000. Modern Mathematical Methods for Physicists and Engineers. Cambridge University Press.

Pearson, K., 1895. Note on Regression and Inheritance in The Case of Two Parents. Proceedings of the Royal Society. 58, (pp 240–242).

Singhal, A., 2001. Modern Information Retrieval A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. 24 (4): 35–43.

Hamming, R. W., 1950, Error Detecting and Error Correcting Codes. Bell System Technical Journal. 29(2): 147–160

Jaccard, P., 1901, Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles. 37: 547–579.

Pongsakorn Teerarassamee

He is currently a master student with the School of Computer Engineering, Suranaree University of Technology, Thailand. His current research of interest includes Classification

Kittisak Kerdprasop

He is an associate professor and chair of the School of Computer Engineering, Suranaree University of Technology, Thailand. His current research of interest includes Data mining, Artificial Intelligence, Functional and Logic Programming Languages, Computational Statistics.

Nittaya Kerdprasop

She is an associate professor at the School of Computer Engineering, Suranaree University of Technology, Thailand. She is a member of ACM and IEEE Computer Society. Her research of interest includes Knowledge Discovery in Databases, Artificial Intelligence, Logic Programming, and Biomedical Informatics

Proceedings of the 3rd International Conference on Industrial Application Engineering 2015

An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm

Kittipong Chomboon*, Pasapitch Chujai, Pongsakorn Teerarassamee, Kittisak Kerdprasop, Nittaya Kerdprasop

School of Computer Engineering, Institute of Engineering, Suranaree University of Technology, Nakhorn Ratchasima 3000, Thailand.

*Corresponding Author: chomboon.k@gmail.com

Abstract

This research aims at studying the performance of k-nearest neighbor classification when applying different distance measurements. In this work, we comparatively study 11 distance metrics including Euclidean, Standardized Euclidean, Mahalanobis, City block, Minkowski, Chebychev, Cosine, Correlation, Hamming, Jaccard, and Spearman. A series of experimentations has been performed on eight synthetic datasets with various kinds of distribution. The distance computations that provide highly accurate prediction consist of City block, Chebychev, Euclidean, Mahalanobis, Minkowski, and Standardize Euclidean techniques.

Keywords: Data Classification, Synthetic Data, Distance Metrics, k-Nearest Neighbors.

1. Introduction

Data mining is the extraction of knowledge hidden in the data. Data mining is often done with the large datasets. The knowledge from data mining has been used in various fields, such as prediction over future situation, assisting in medical diagnosis, forecasting relation of chronology.

Current data mining methodology has been classified into several tasks, such as classification, clustering, and association mining. Data mining each for task will have a different purpose. Classification task will be trying to classify data with high accuracy for classifying future example, such as trying to distinguish between patients with heart disease and those who are healthy. Clustering task will try to categorize groups of data such that data in the same group look similar, whereas they are dissimilar to others in different groups. Association mining task will try to find rules that represent relation between data with some support and confident values.

Classification task of data mining can be done with many algorithms such as k-nearest neighbor. Beyer[1] explained the significance and origin of the nearest neighbor. Cover[2] used k-nearest neighbor to classify data. Dudani[3] did research about weighting of distance matrix values with k-nearest neighbor. Fukunaga[4] developed techniques for running k-nearest neighbor faster. Keller[5] developed new algorithm named “Fuzzy K-Nearest Neighbor” based on k-nearest neighbor with the purpose to use it with fuzzy task. Köhn[6] used city-block distance matric to increase performance of k-nearest neighbor algorithm.

This research also studies classification technique with a specific interest in the k-nearest algorithm. We aim to analyze the performance of different distance metrics to finally choose a proper metric that makes a good classification performance. In this research use 8 synthetic datasets with different distribution, and a dataset for each distribution has 2 classes but has different amount of data in each class. This is to test the impact about amount in each class on the performance of classification.

The rest of this research is organized as follows: Section 2 gives details of the k-Nearest Neighbor and the computation of each distance metric. Section 3 gives details of our proposed method. The experimental results and analysis will be presented in Section 4. Finally, the research is concluded in Section 5.

DOI: 10.12792/iciae2015.051

© 2015 The Institute of Industrial Applications Engineers, Japan.

2. Background

2.1 k-Nearest Neighbor

The k-nearest neighbor is a semi-supervised learning algorithm such that it requires training data and a predefined k value to find the k nearest data based on distance computation. If k data have different classes, the algorithm predicts class of the unknown data to be the same as the majority class. For example, to find the appropriate class of new datum using the k-nearest neighbor algorithm with a Euclidean distance metric, the concept can be shown in Fig. 1.

Fig. 1 shows the classification of iris data. The point to be classified is (5, 1.45), which is shown with "X". When applying k-nearest neighbor algorithm with k = 8 using Euclidean distance computation, the result is shown with a radius of dot line. It has two possible classes: virginica class with two instances and versicolor class with six instances. This algorithm will classify mark "X" to the class of versicolor because versicolor class is the majority of data within the radius.

2.2 Distance Metrics

Distance metrics are a method to find distance between a new data point and existing training dataset. In this research, we experiment with 11 distance metrics, which can be explained as follows.

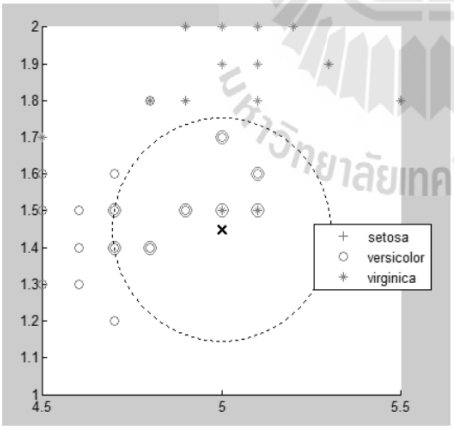


Fig. 1. The k-nearest neighbor prediction with k = 8.

Given an *mx*-by-*n* data matrix X, which is treated as *mx* (1-by-*n*) row vectors x_1, x_2, ..., xm_x, and *my*-by-*n* data matrix Y, which is treated as *my*(1-by-*n*) row vectors y_1, y_2,...,ym_y, the various distances between the vectors x_s and y_t are defined as follows:

1. Euclidean Distance

The Euclidean distance is a measure to find distance between two points, defined by Eq. (1)

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)' \quad (1)$$

The Euclidean distance is a special case of the Minkowski metric, where p = 2.

2. Standardized Euclidean Distance

The standardized Euclidean distance is used to optimize the problem of finding the distance, defined by Eq. (2)

$$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)' \quad (2)$$

where V is the n-by-n diagonal matrix whose jth diagonal element is $S(j)^2$, S is the vector containing the inverse weights.

3. Mahalanobis Distance

The Mahalanobis distance is a measure between a point and a distribution of data, defined by Eq. (3)

$$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)' \quad (3)$$

where C is the covariance matrix.

4. City Block Distance

The city block distance between two points is the sum of the absolute difference of Cartesian coordinates, defined by Eq. (4)

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \quad (4)$$

The city block distance is a special case of the Minkowski metric, where p = 1.

5. Minkowski Distance

The Minkowski distance is a method to find distance based on Euclidean space, defined by Eq. (5)

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - y_{tj}|^p} \quad (5)$$

For the special case of Minkowski distance

p = 1, the Minkowski metric gives the city block distance,

p = 2, the Minkowski metric gives the Euclidean distance, and

p = ∞, the Minkowski metric gives the Chebychev distance.

6. Chebychev Distance

The Chebychev distance is a measure to find distance between two vectors or points with standard coordinates, defined by Eq. (6)

$$d_{st} = max_j\{|x_{sj} - y_{tj}|\} \quad (6)$$

The Chebychev distance is a special case of the Minkowski metric, where p = ∞.

7. Cosine Distance

The Cosine distance is computed from one minus the cosine of the included angle between points, defined by Eq. (7)

$$d_{st} = \left(1 - \frac{x_s {y'}_t}{\sqrt{(x_s {x'}_s)(y_t {y'}_t)}}\right) \quad (7)$$

8. Correlation Distance

Distance based on correlation is a measure of statistical dependence between two vectors, defined by Eq. (8)

$$d_{st} = \left(1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}}\right) \quad (8)$$

where

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$$

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj}$$

9. Hamming Distance

Hamming distance, which is the percentage of coordinates that differ, can be defined by Eq. (9)

$$d_{st} = \left(\frac{\#(x_{sj} \neq y_{tj})}{n}\right) \quad (9)$$

10. Jaccard Distance

Jaccard distance is computed from one minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ, defined by Eq. (10)

$$d_{st} = \left(\frac{\#\left[(x_{sj} \neq y_{tj}) \cap \left((x_{sj} \neq 0) \cup (y_{tj} \neq 0)\right)\right]}{\#\left[(x_{sj} \neq 0) \cup (y_{tj} \neq 0)\right]}\right) \quad (10)$$

11. Spearman Distance

Spearman distance is computed from one minus the sample Spearman's ranked correlation between observations, defined by Eq. (11)

$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}} \quad (11)$$

Where

r_{sj} is the rank of x_{sj} taken over x_1j, x_2j, ...xmx,j.

r_{tj} is the rank of y_{tj} taken over y_1j, y_2j, ...ymy,j.

r_s and r_t are the coordinate-wise rank vectors of x_s and y_t,

i.e., $r_s = (r_{s1}, r_{s2}, \ldots r_{sn})$ and $r_t = (r_{t1}, r_{t2}, \ldots r_{tn})$.

$$\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{(n+1)}{2}$$

$$\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{(n+1)}{2}$$

3. Empirical Study Methodology

In this section, we present our study framework using k-nearest neighbor algorithm with various distance metrics. The framework is shown in Fig. 2.

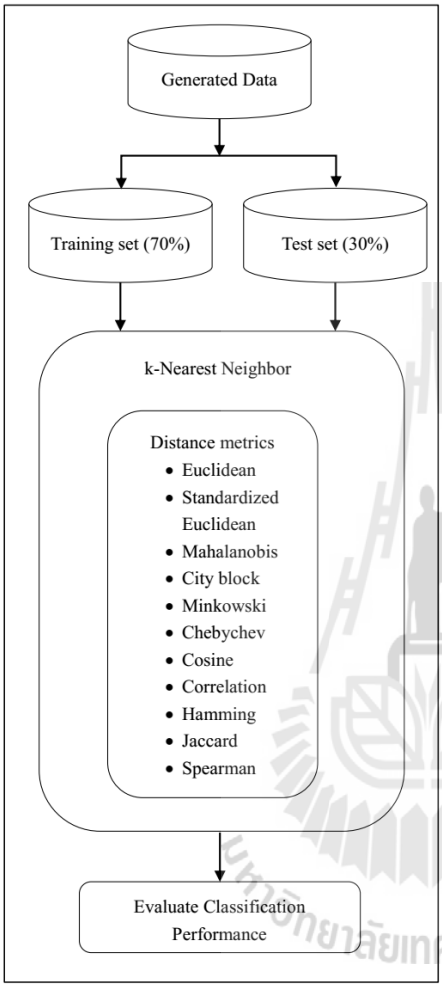


Fig. 2. The framework of our empirical study.

From Fig. 2 the detail of each step can be explained as follows:

Step 1: Generate binary data set with different distribution and different amount of data in each class. Then split data around 70% for training set and 30% for test set, which will be used for testing the performance of classification.

Step 2: Use data from step 1 for data classification by applying the k-nearest neighbor algorithm with various distance metrics to compute the k-nearest data points for making classification.

Step 3: Analyze the results and conclude about the performance of classification using various distance metrics.

4. Experimental Results

4.1 Datasets

For our experiment, the proposed framework has been applied for classifying binary synthetic datasets. We generate eight synthetic datasets, each dataset has four different distributions, and each distribution has two of data in which class the amount of data in each class is varied. Each dataset has in total 5000 instances, and three features. We use MATLAB program to generate synthetic datasets. Details of the synthetic datasets are given in Table 1. Fig. 3 illustrates an overview of synthetic datasets.

Table 1. Details of synthetic datasets.

Dataset	Mean	SD	Class 1	Class 2	Total
1	[0 0 0; 3 0 0]	[1 0 0; 0 1 0; 0 0 1]	2500	2500	5000
2	[0 0 0; 3 0 0]	[1 0 0; 0 1 0; 0 0 1]	4750	250	5000
3	[0 0 0; 0 0 3]	[0.2 0 0; 0 0.2 0; 0 0 1]	2500	2500	5000
4	[0 0 0; 0 0 3]	[0.2 0 0; 0 0.2 0; 0 0 1]	4750	250	5000
5	[0 0 0; 3 0 0]	[1 0 0; 0 0.2 0; 0 0 0.2]	2500	2500	5000
6	[0 0 0; 3 0 0]	[1 0 0; 0 0.2 0; 0 0 0.2]	4750	250	5000
7	[0 0 0; 3 3 0]	[1 0.9 0; 0.9 1 0; 0 0 1]	2500	2500	5000
8	[0 0 0; 3 3 0]	[1 0.9 0; 0.9 1 0; 0 0 1]	4750	250	5000

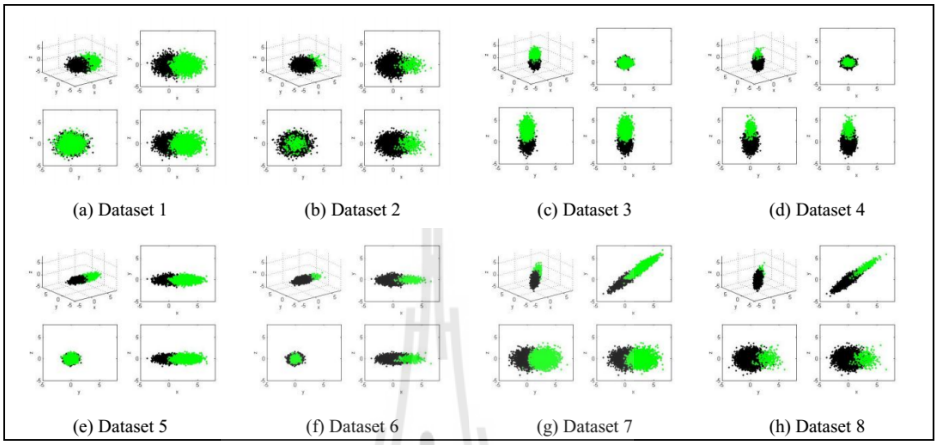


Fig. 3. Distribution of the eight synthetic datasets, each one has four kinds of distribution.

4.2 Experimental Results

The results from the proposed study framework for eight synthetic datasets have been shown in Figs. 4 and 5. The data classification has been performed with the same algorithm (that is, k-Nearest Neighbor) and the same parameter setting. The only varied factor is a distance measurement. It turns out that the Hamming and Jaccard distance metrics perform badly on 4 out of 8 datasets.

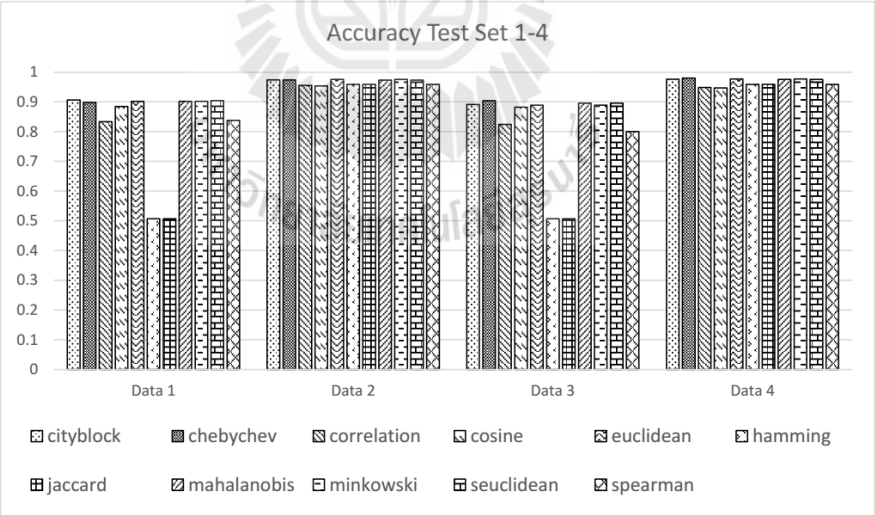


Fig. 4. Accuracy of synthetic datasets from no. 1 to no. 4.

 © 2015 The Institute of Industrial Applications Engineers, Japan.

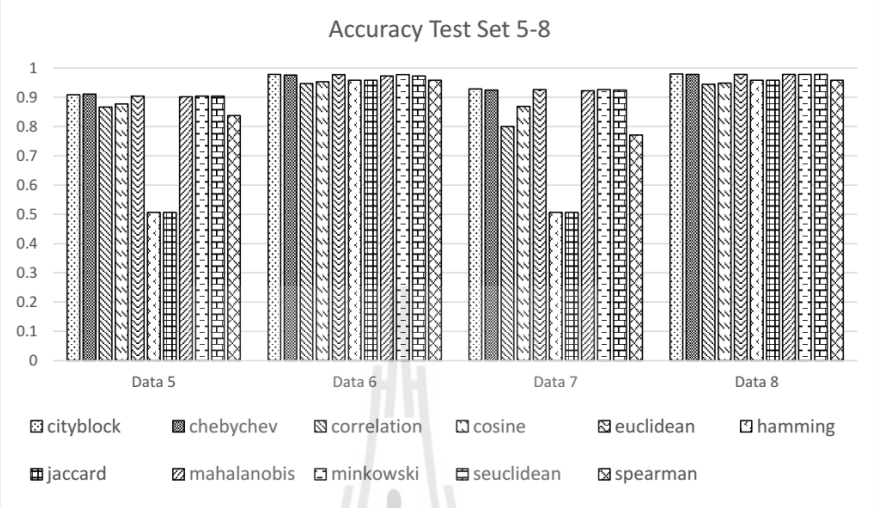


Fig. 5. Accuracy of synthetic datasets from no. 5 to no. 8.

5. Conclusions

The results of this research showed accuracy of k-nearest neighbor classification algorithm with different distance metrics. Experiments had been performed on eight synthetic datasets generated by MATLAB. The synthetic datasets have four distributions and have been split 70% to training set and 30% to test set. The results of classification over datasets in which amount of data in each class is equal showed that the Hamming and Jaccard techniques are low accuracy, while the other distance computation techniques have similar accuracy. The synthetic datasets in which amount of data in each class is different such as dataset 2, 4, 6 and 8 showed that the Hamming and Jaccard techniques are increasing in their classification accuracy. We can conclude that Hamming and Jaccard techniques are affected by the ratio of members in each class, while the other techniques are not affected by such phenomenon. The highest accuracy on classify data with k-Nearest Neighbor is obtained from the six distance metrics, that are City-block, Chebychev, Euclidean, Mahalanobis, Minkowski, and Standardized Euclidean techniques.

References

(1) Beyer Kevin, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft, When is “nearest neighbor” meaningful?, in Database Theory—ICDT’99. 1999, Springer. p. 217-235.aaaa

(2) Cover Thomas and Peter Hart: "Nearest neighbor pattern classification", Information Theory, IEEE Transactions on, Vol. 13, No. 1, pp. 21-27, 1967

(3) Dudani Sahibsingh A: "The distance-weighted k-nearest-neighbor rule", Systems, Man and Cybernetics, IEEE Transactions on, Vol. No. 4, pp. 325-327, 1976

(4) Fukunaga Keinosuke and Patrenahalli M Narendra: "A branch and bound algorithm for computing k-nearest neighbors", Computers, IEEE Transactions on, Vol. 100, No. 7, pp. 750-753, 1975

(5) Keller James M, Michael R Gray, and James A Givens: "A fuzzy k-nearest neighbor algorithm", Systems, Man and Cybernetics, IEEE Transactions on, Vol. No. 4, pp. 580-585, 1985

(6) Köhn Hans-Friedrich: "Combinatorial individual differences scaling within the city-block metric", Computational Statistics & Data Analysis, Vol. 51, No. 2, pp. 931-946, 2006

ประวัติผู้เขียน

นายพงศกร ธีรรัศมี เกิดเมื่อวันที่ 23 มกราคม พ.ศ. 2535 ภูมิลำเนาเดิม จังหวัดหนองคาย เริ่มเข้าศึกษาระดับชั้นอนุบาล 1 ถึงอนุบาล 3 ที่โรงเรียนนิคมศึกษา ชั้นประถมศึกษาปีที่ 1 ถึงชั้นประถมศึกษาปีที่ 6 ที่โรงเรียนฮั่วเคียวกงฮัก จากนั้นได้เข้าศึกษาต่อในระดับมัธยมศึกษาตอนต้นที่โรงเรียนโรซารีโอวิทยา ในระดับมัธยมศึกษาตอนปลายที่โรงเรียนปทุมเทพวิทยาคาร อำเภอเมือง จังหวัดหนองคาย และได้เข้าศึกษาต่อระดับปริญญาตรีในสาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรานารี อำเภอเมือง จังหวัดนครราชสีมา และสำเร็จการศึกษาเมื่อปี พ.ศ. 2556 ภายหลังสำเร็จการศึกษาในระดับปริญญาตรี ได้เข้าศึกษาในระดับปริญญาโท สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี ในปี 2557

ในระหว่างการศึกษาได้รับความอนุเคราะห์อย่างดียิ่งจากอาจารย์ประจำวิชา Database System ได้รับความไว้วางใจให้เป็นผู้ช่วยสอนปฏิบัติการ และได้รับการตีพิมพ์เผยแพร่บทความวิชาการซึ่งรายละเอียดสามารถดูได้ที่ภาคผนวก ค

มหาวิทยาลัยเทคโนโลยีสุรนารี