การจำแนกข้อมูลโดยการเลือกคุณลักษณะที่สำคัญ

DATA CLASSIFICATION BY SELECTING IMPORTANT ATTRIBUTES

จิราภรณ์ เตชะแก้ว

JIRAPORN TEOKAEW

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2555

KMITL-2012-SC-M-028-007

การจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่สำคัญ

DATA CLASSIFICATION BY SELECTING IMPORTANT ATTRIBUTES

จิราภรณ์ ถมแก้ว

JIRAPORN THOMKAEW

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2555

KMITL–2012–SC-M-002-007

DATA CLASSIFICATION BY SELECTING IMPORTANT ATTRIBUTES

JIRAPORN THOMKAEW

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2012

KMITL–2012–SC-M-002-007

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์	การจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่สำคัญ Data Classification by Selecting Important Attributes
นักศึกษา	นางสาวจิราภรณ์ ถมแก้ว
รหัสประจำตัว	52650809
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร.ศรัณย์ อินทโกสุม

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
รศ.ดร.จีรพร	วีระพันธุ์	
รศ.ดร.วีระ	บุญจริง	
ดร.เฉลิมศักดิ์	เลิศวงศ์เสถียร	
ผศ.ดร.ศรัณย์	อินทโกสุม	

วัน / เดือน/ ปี ที่สอบ 20 เมษายน พ.ศ. 2555 เวลา 9.00 – 12.00 น.
สถานที่สอบ ณ ห้อง 216 ชั้น 2 อาคารจุฬาภรณวลัยลักษณ์ 1

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.คุษณี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์

วันที่ 26 **เดือน** เม.ย. **พ.ศ.** 55

หัวข้อวิทยานิพนธ์	การจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่สำคัญ
นักศึกษา	นางสาวจิราภรณ์ ถมแก้ว
รหัสประจำตัว	52650809
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2555
อาจารย์ที่ปรึกษา	ผศ.ดร.ศรัณย์ อินทโกสุม

บทคัดย่อ

การจำแนกข้อมูลจากงานวิจัยที่ผ่านมามักจะพิจารณาคุณลักษณะทั้งหมดของชุดข้อมูล อย่างไรก็ตามคุณลักษณะบางประการมีความสำคัญน้อยซึ่งเมื่อนำมาร่วมคำนวณด้วยแล้วอาจเป็นสาเหตุทำให้ความแม่นยำในการจำแนกข้อมูลลดลง งานวิจัยนี้เสนอวิธีการใหม่เพื่อปรับปรุงประสิทธิภาพการจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่ไม่สำคัญของชุดข้อมูลก่อนการจำแนกข้อมูล วิธีนี้ได้พัฒนาบนพื้นฐานของการทดลองดังนี้ ขั้นตอนแรกทดลองเพื่อเลือกอัลกอริธึมสำหรับการจำแนกประเภทข้อมูลที่เหมาะสม โดยการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยสามอัลกอริธึม คือ Radial Basis Function, Multi–Layer Perceptron และ Naïve Bays ผลการทดลองพบว่า จำแนกข้อมูลด้วย Multi - Layer Perceptron ให้ค่าความแม่นยำในการจำแนกประเภทข้อมูลสูงแต่ใช้เวลาในการประมวลผลมาก และการจำแนกข้อมูลด้วย Naïve Bays ใช้เวลาในการประมวลผลน้อยแต่ให้ค่าความแม่นยำในการจำแนกข้อมูลค่อนข้างต่ำ การจำแนกข้อมูลด้วย Radial Basis Function (RBF) ให้ผลลัพธ์ที่ยอมรับได้ในทั้งสองด้าน คือให้ค่าความแม่นยำในระดับที่ยอมรับได้และใช้เวลาในการประมวลผลที่เหมาะสม ดังนั้นจึงเลือกใช้ RBF สำหรับการจำแนกประเภทข้อมูล ขั้นตอนที่สองเป็นการทดลองเพื่อหาอัลกอริธึมที่ดีที่สุดสำหรับการคัดเลือกคุณลักษณะที่ไม่สำคัญของชุดข้อมูล ขั้นตอนนี้ทดลองโดยใช้สามอัลกอริธึมคือ Greedy Algorithm, Information Gain และ Principal Component Analysis ผลการทดลองพบว่าการใช้อัลกอริธึมแต่ละตัวก่อนที่จะใช้ RBF สามารถเพิ่มระดับค่าความแม่นยำของการจำแนกข้อมูลได้ดีกว่าการใช้ RBF เพียงอย่างเดียว อย่างไรก็ตามเมื่อเปรียบเทียบกันทั้งสามอัลกอริธึมแล้วพบว่า Greedy Algorithms ให้ประสิทธิภาพดีที่สุด จึงสามารถสรุปได้ว่าถ้าต้องการได้ค่าความแม่นยำของการจำแนกข้อมูลดีที่สุด ควรใช้ Greedy Algorithms ในการคัดเลือกคุณลักษณะที่ไม่สำคัญของข้อมูลก่อนแล้วจึงจำแนกข้อมูลด้วย RBF

คำสำคัญ: การคัดเลือกคุณลักษณะ, เรเดียลเบสิสฟังก์ชัน, กรีดดีอัลกอริธึม

Thesis Title	Data Classification by Selecting Important Attributes
Student	Jiraporn Thomkaew
Student ID	52650809
Degree	Master of Science
Program	Computer Science
Year	2012
Thesis Advisor	Asst.Prof.Dr.Sarun Intakosum

ABSTRACT

The existing researches in data classification always consider all data attributes. However, there may be some attributes that are unimportant and may reduce the accuracy ratio of data classification. This research proposes a new method to improve the data classification performance by filtering out unimportant data attributes before performing the data classification process. The proposed method has been developed based on the following experiments. The first experiment aims to choose the proper data classification algorithms. This has been done by comparing the performance of the following three data classification algorithms; Radial Basis Function, Multi–Layer Perceptron, and Naïve Bays. The results have shown that Multilayer-Perceptron gives the highest accuracy classification rate but requires much processing time. On the other hands, Naïve Bays uses less processing time but provides poor accuracy. Radial Basis Function (RBF) is the compromise between both since it gives the acceptance accuracy level with the proper processing time. Therefore the RBF has been chosen to use in the second experiment that aims to find the best algorithm to filter out unimportant data attributes. In this step, three algorithms namely Greedy Algorithm, Information Gain, and Component Analysis have been applied. The results have shown that applying each algorithm before RBF can improve the accuracy level of the data classification process that has been done previously using RBF alone. However, among the three algorithms, Greedy Algorithm gives the best performance. This can be concluded that in order to gain the best data classification accuracy rate, the Greedy Algorithm should be used to select out unimportant attributes before applying RBF to perform the data classification process.

Keywords: Attributes Selecting, Radial Basis Function, Greedy Algorithms

กิตติกรรมประกาศ

วิทยานิพนธ์นี้มิอาจสำเร็จลุล่วงไปได้ด้วยดี หากมิได้รับคำแนะนำ คำชี้แจง ความรู้และความเอาใจใส่จาก ผศ.ดร.ศรัณย์ อินทโกสุม ผู้เป็นอาจารย์ที่ปรึกษา ซึ่งท่านได้สละเวลาให้กับข้าพเจ้าอย่างเต็มที่ จึงใคร่ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ รศ.ดร.จีรพร วีระพันธุ์ รศ.ดร.วีระ บุญจริง และ ดร.เฉลิมศักดิ์ เลิศวงษ์เสถียร คณะกรรมการสอบหัวข้อ และโครงร่างวิทยานิพนธ์ ที่กรุณาให้คำแนะนำตลอดจนข้อชี้แนะจนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

ขอขอบพระคุณบิดา มารดา และครอบครัว ที่สนับสนุนให้ได้เรียนในระดับที่ตั้งใจ และให้ความรัก ความอบอุ่น ความเข้าใจ อีกทั้งยังดูแลเรื่องค่าใช่จ่ายต่างๆ ระหว่างศึกษาเป็นอย่างดีอีกด้วย

ขอขอบคุณพี่ๆ น้องๆ และเพื่อนๆ ทุกคนที่ให้คำปรึกษา และช่วยอำนวยความสะดวกในด้านต่างๆ

สำหรับคุณงามความดีและประโยชน์อันใดที่เกิดขึ้นจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดา มารดา อาจารย์ทุกท่านซึ่งเป็นที่เคารพรักยิ่ง ตลอดจนญาติพี่น้อง และเพื่อนๆ ทุกคน

จิราภรณ์ ถมแก้ว

เมษายน 2555

สารบัญ

หน้า

สารบัญ(ต่อ)

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

รูปที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การจำแนกข้อมูล เป็นเทคนิคหนึ่งในการทำเหมืองข้อมูล โดยการจำแนกข้อมูลให้อยู่ในกลุ่มเดียวกันตามที่กำหนดโดยการสร้างกฎเพื่อช่วยการตัดสินใจจากข้อมูลที่มีอยู่ และมีการนำมาประยุกต์ใช้กับงานด้านต่างๆ เช่น การจำแนกผู้ป่วยโรคต่างๆ การจำแนกกลุ่มลูกค้า การจำแนกเอกสารสารสนเทศ เป็นต้น ซึ่งการจำแนกข้อมูลดังกล่าวต้องใช้คุณลักษณะของข้อมูลเป็นสำคัญ

คุณลักษณะของข้อมูล คือ ลักษณะหรือคุณสมบัติที่ใช้ระบุองค์ประกอบหรือรายละเอียดของชุดข้อมูล คุณลักษณะของข้อมูลที่ดีต้องมีความถูกต้องและเชื่อถือได้ เก็บเฉพาะข้อมูลที่จำเป็นต้องใช้และครบถ้วนสมบูรณ์ อย่างไรก็ตามหากชุดข้อมูลมีการเก็บคุณลักษณะที่มากเกินความจำเป็นจะทำให้สิ้นเปลืองทรัพยากรและเวลา การนำข้อมูลไปใช้งานอาจทำให้ประสิทธิภาพลดลงได้ การคัดเลือกคุณลักษณะของชุดข้อมูล ก่อนจำแนกข้อมูลเป็นกระบวนการเพื่อลดขนาดมิติของข้อมูลเดิม [9] แต่ทำให้สูญเสียความสำคัญของข้อมูลน้อยที่สุด

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

งานวิจัยนี้มุ่งศึกษาอัลกอริธึมสำหรับการคัดเลือกคุณลักษณะที่สำคัญของชุดข้อมูลก่อนการจำแนกข้อมูล เพื่อเพิ่มค่าความแม่นยำในการจำแนกข้อมูลให้สูงขึ้นและช่วยลดเวลาในการทำงาน

1.3 สมมติฐานของการศึกษา

การใช้กรีดดีอัลกอริธึมสำหรับคัดเลือกคุณลักษณะที่สำคัญของชุดข้อมูล ก่อนการจำแนกข้อมูลด้วย Radial Basis Function จะสามารถเพิ่มค่าความแม่นยำสูงกว่าการจำแนกข้อมูลโดยใช้คุณลักษณะทั้งหมดของชุดข้อมูล

1.4 ขอบเขตการศึกษา

วิทยานิพนธ์นี้มีขอบเขตของการวิจัย ดังนี้

1. ศึกษาและเลือกชุดข้อมูลสำหรับการทดสอบจาก UCI Machine Learning Repository Data Sets [10]

2. วิเคราะห์วิธีการและความสามารถในการจำแนกข้อมูลด้วย Radial basis Function แบบใช้คุณลักษณะทั้งหมดของชุดข้อมูล

3. ใช้เครื่องมือในการประมวลผลด้วยโปรแกรม Weka 3.6 [11]

4. งานวิจัยนี้พัฒนาวิธีการเพิ่มค่าความแม่นยำในการจำแนกข้อมูลด้วยวิธีการคัดเลือกคุณลักษณะที่สำคัญของชุดข้อมูลโดยการนำเทคนิคของกรีดดีอัลกอริธึมมาประยุกต์ใช้

1.5 ขั้นตอนการศึกษาและดำเนินงานวิจัย

วิทยานิพนธ์นี้มีขั้นตอนการศึกษาและดำเนินงานวิจัย ดังนี้

1. ศึกษาวิเคราะห์การจำแนกข้อมูลแบบใช้คุณลักษณะทั้งหมดของชุดข้อมูล

2. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

3. ตั้งสมมติฐาน โดยคาดว่าการใช้กรีดดีอัลกอริธึมคัดเลือกคุณลักษณะที่สำคัญของชุดข้อมูลก่อนการจำแนกข้อมูลด้วย Radial Basis Function จะให้ค่าความแม่นยำสูงกว่าการใช้คุณลักษณะทั้งหมดในการจำแนกข้อมูล

4. ทดสอบตามกระบวนการที่นำเสนอ

5. วิเคราะห์ผลการทดลอง

6. สรุปผลการทดลองและเสนอแนวทางการพัฒนางานวิจัยครั้งต่อไป

7. เขียนวิทยานิพนธ์

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการคัดเลือกคุณลักษณะที่สำคัญของชุดข้อมูล และการจำแนกประเภทข้อมูล โดยในส่วนที่หนึ่งจะกล่าวถึงทฤษฎีและงานวิจัยเกี่ยวกับการคัดเลือกคุณลักษณะของข้อมูล ส่วนที่สองจะกล่าวถึงทฤษฎีและงานวิจัยเกี่ยวกับการจำแนกประเภทข้อมูล

2.1 การคัดเลือกคุณลักษณะ

คุณลักษณะของข้อมูล คือ ลักษณะหรือคุณสมบัติที่ใช้ระบุองค์ประกอบหรือรายละเอียดของชุดข้อมูล คุณลักษณะของข้อมูลที่ดีต้องมีความถูกต้องและเชื่อถือได้ เก็บเฉพาะข้อมูลที่จำเป็นต้องใช้และครบถ้วนสมบูรณ์ อย่างไรก็ตามหากชุดข้อมูลมีการเก็บคุณลักษณะที่มากเกินความจำเป็นจะทำให้สิ้นเปลืองทรัพยากรและเวลา การนำข้อมูลไปใช้งานอาจทำให้ประสิทธิภาพลดลงได้

การคัดเลือกคุณลักษณะ คือการลดขนาดของข้อมูลโดย [9] หลักการที่สำคัญของการลดขนาดข้อมูลคือ การทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด เนื่องจากข้อมูลแต่ละตัวจะมีความสำคัญต่อการจัดกลุ่มข้อมูลไม่เท่ากัน ด้วยเทคนิคการเลือกข้อมูลที่ดีจะ ทำให้สามารถเลือกข้อมูลที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ ข้อมูลที่มีการรวมกลุ่มกันอย่างหนาแน่นจะเป็นข้อมูลที่มีความสำคัญต่อการจัดกลุ่มข้อมูลในอนาคต ทฤษฎีการคัดเลือกคุณลักษณะของชุดข้อมูล มีงานวิจัยที่เกี่ยวกับการคัดเลือกคุณลักษณะดังนี้

1.พลอยพรรณ สอนสุวิทย์ [4] นำเสนอวิธีตรวจจับการบุกรุกเครือข่ายชนิด Probing ที่มีคุณลักษณะจำนวนมาก จึงใช้ Genetic Algorithm ในการคัดเลือกคุณลักษณะจากการทดลองพบว่าการคัดเลือกคุณลักษณะที่ไม่สำคัญออก ไม่มีผลต่อค่าความถูกต้องของข้อมูลและยังสามารถลดเวลาการทำงานได้ด้วย

2.วงกต ศรีอุไร [5] ปรับปรุงการจำแนกหมวดหมู่ของเอกสารโดยวิธีสร้างแบบจำลองหัวข้อให้กับเอกสารและใช้วิธีการคัดเลือกคุณลักษณะสองวิธี คือ Information Gain และ Chi Squared จากการทดลองพบว่าการคัดเลือกคุณลักษณะด้วยวิธี Information Gain ให้ประสิทธิภาพดีกว่า มีความแม่นยำเพิ่มขึ้น 10.2%

3.จิราพร สุดใหญ่ [10] นำเสนอการคัดเลือกคุณลักษณะข้อมูลเพื่อลดมิติของตัวแปรที่ไม่มีความสัมพันธ์กัน ขั้นตอนวิธีที่นำมาใช้เพื่อลดมิติข้อมูลคือ Principal Component Analysis, Linear Discriminate Analysis, Maximum margin criterion โดยเลือกตัดมิติค่าความสมนัยที่มีค่าน้อย ผลการทดลองพบว่าการคัดเลือกด้วย Linear Discriminate Analysis ให้ค่าความแม่นยำที่สูงกว่า

การคัดเลือกคุณลักษณะด้วยเทคนิคที่ต่างกันจะทำให้ได้คุณลักษณะที่สำคัญต่างกันด้วย นอกจากนี้ยังมีทฤษฎีอื่นๆ ที่ใช้สำหรับการคัดเลือกคุณลักษณะ เช่น

2.1.1 Rough Set

เป็นทฤษฎีที่ถูกนำเสนอโดย Zdzislaw Pawlak ในปี ค.ศ. 1982 เป็นทฤษฎีคณิตศาสตร์แนวใหม่ ที่ใช้ในการจัดการเกี่ยวกับเรื่องความคลุมเครือและความไม่แน่นอนของข้อมูลซึ่งทฤษฎีนี้จะอิงพื้นฐานในเรื่องเกี่ยวกับปัญญาประดิษฐ์ (AI: Artificial Intelligence) เป็นสำคัญและเป็นทฤษฎีที่สัมพันธ์และอยู่ในขอบเขตแนวทางเดียวกันกับการศึกษาเกี่ยวกับเรื่องการเรียนรู้ (Machine learning) การวิเคราะห์การตัดสินใจ (Decision Analysis) การค้นหาความรู้ที่ต้องการจากฐานข้อมูล (Knowledge Discovery from Database) ระบบผู้เชี่ยวชาญ (Expert System) และอื่นๆ

ลักษณะของข้อมูลที่มีความคลุมเครือและขัดแย้งกันของข้อมูล ทฤษฎี Rough Set สามารถช่วยลดความคลุมเครือของข้อมูลได้ ซึ่งจะทำให้ได้ผลลัพธ์ 2 อย่าง คือ ผลลัพธ์ที่หนึ่งเรียกว่า การประมาณขอบเขตล่าง (R-lower approximation) ซึ่งจะได้กลุ่มตัวอย่างของข้อมูลที่ไม่คลุมเครือแน่นอน และ ผลลัพธ์ที่สองเรียกว่า การประมาณขอบเขตบน (R-upper approximation) ซึ่งจะได้กลุ่มตัวอย่างข้อมูลที่มีความเป็นไปได้ว่าจะเป็นผลลัพธ์ที่หนึ่ง

ทฤษฎี Rough Set จะใช้กับข้อมูลที่มีความคลุมเครือ และความไม่แน่นอน ซึ่งอยู่บนสมมติฐานที่ว่า สามารถหาฐานความรู้ของข้อมูล จากทุกวัตถุในเอกภพสัมพันธ์ วิธีการทำงาน คือ

ลดคุณสมบัติที่ไม่จำเป็นและหาค่าความสำคัญของคุณลักษณะที่ดีที่สุดเพื่อให้ได้กฎความรู้สำหรับนำไปใช้งาน

2.1.2 Information Gain

เป็นเทคนิคการคัดเลือกคุณลักษณะของข้อมูลเพื่อลดมิติของข้อมูล โดยการวัดค่าความสำคัญของแต่ละโหนด ถ้าโหนดใดมีค่าความสำคัญสูงสุดก็จะถูกเลือกเป็นโหนดราก และนำข้อมูลที่เหลือมาหาค่าความสำคัญอีกครั้งเพื่อให้ได้โหนดต่อไป นิยมใช้กับลักษณะของข้อมูลที่เป็นแบบไม่ต่อเนื่อง ซึ่งคำนวณได้จาก

$$Gain(S,A) = E(S) - \sum_{v \in V(A)} \frac{S_v}{S} E(S_v) \qquad (2.1)$$

เมื่อ Gain(S,A) = ค่า Gain ของเหตุการณ์ที่สนใจ

$E(S_v)$ = ค่า Entrophy ของเซ็ตข้อมูลที่สนใจ

S_v = เซ็ตข้อมูลที่สนใจ

ซึ่งค่า Entrophy สามารถคำนวณได้จาก

$$E(S) - \sum_{i=1}^{n} - P(V_i) \log_2 P(V_i) \qquad (2.2)$$

เมื่อ E(S) = ค่า Entrophy ของเซ็ตข้อมูลทั้งหมด

$S = P(V_1), P(V_2), \ldots, P(V_n)$

$P(V_i)$ = ค่าความน่าจะเป็นของข้อมูลที่สนใจ

2.1.3 Greedy Algorithm

เป็นการค้นหาแบบดีที่สุดก่อน (Best First Search) โดยพิจารณาข้อมูลที่มีอยู่ในขณะนั้น มีทางเลือกใดที่ให้ผลตอบแทนคุ้มที่สุด วิธีนี้จะหาทางเลือกที่ดีที่สุดในขณะนั้น โดยการสร้างกราฟต้นไม้ จากนั้นจะค้นหาแบบชั้นต่อชั้น ระหว่างโหนดจะเชื่อมต่อกันด้วยกิ่งที่มีค่าน้ำหนักของแต่ละกิ่ง โดยทั่วไปจะใช้ Greedy algorithm กับปัญหาที่ต้องการคำตอบที่เหมาะสมที่สุด (Optimization problem) เพราะต้องการการตัดสินใจว่าทางเลือกในปัจจุบันมีค่าตอบแทนมากที่สุดหรือน้อยที่สุดหรือไม่ คำนวณได้จาก

Input: G = (V;E) (2.3)

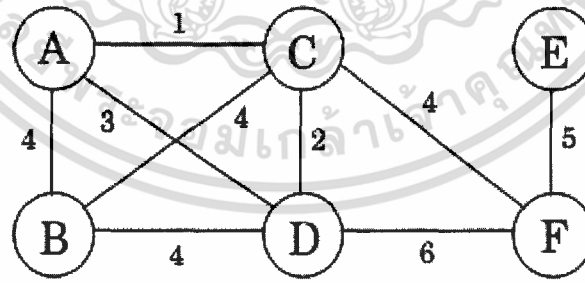
เมื่อ V = จำนวนโหนด, E = จำนวนกิ่ง

Output: T = (V; E'), เมื่อ $E' \subseteq E$ (2.4)

Weight (T) = $\sum w_e$ (2.5)

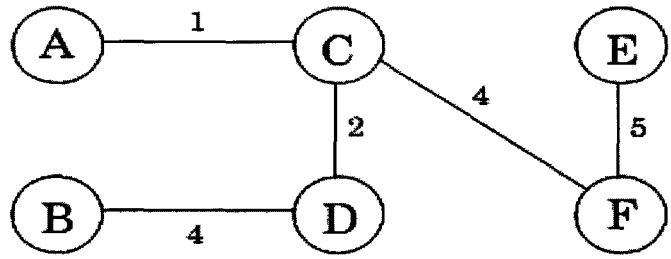
เมื่อ w_e = น้ำหนักแต่ละกิ่ง

เทคนิคของกรีดดีอัลกอริธึม จะพิจารณาเลือกทางเลือกที่สามารถเชื่อมต่อกันได้ทุกโหนด แต่ไม่ก่อให้เกิดเป็นกราฟวงกลม และมีค่าน้ำหนักรวมของทุกโหนดน้อยที่สุด



รูปที่ 2.1 แสดงทางเลือกที่เป็นไปได้ทั้งหมดจากการทำงานของกรีดดีอัลกอริธึม

จากรูปที่ 2.1 แสดงทางเลือกที่เป็นได้ทั้งหมดและค่าน้ำหนักประจำกิ่ง ซึ่งจะนำมาคำนวณหาทางเลือกที่ดีที่สุดด้วยกรีดดีอัลกอริธึม ทางเลือกที่ได้ดังรูปที่ 2.2



รูปที่ 2.2 แสดงทางเลือกที่ได้จากการทำงานของกรีดดีอัลกอริธึม

จากรูปที่ 2.2 แสดงทางเลือกที่ได้จากการทำงานของกรีดดีอัลกอริธึม ซึ่งมีผลรวมของค่าน้ำหนักเท่ากับ 16 อย่างไรก็ตาม ทางเลือกนี้คือทางเลือกที่เป็นไปได้แต่อาจจะไม่ใช่ทางเลือกที่ดีที่สุด ดังนั้นหากจะนำอัลกอริธึมนี้ไปใช้ในการทดลองควรพิจารณาแต่ละกรณี

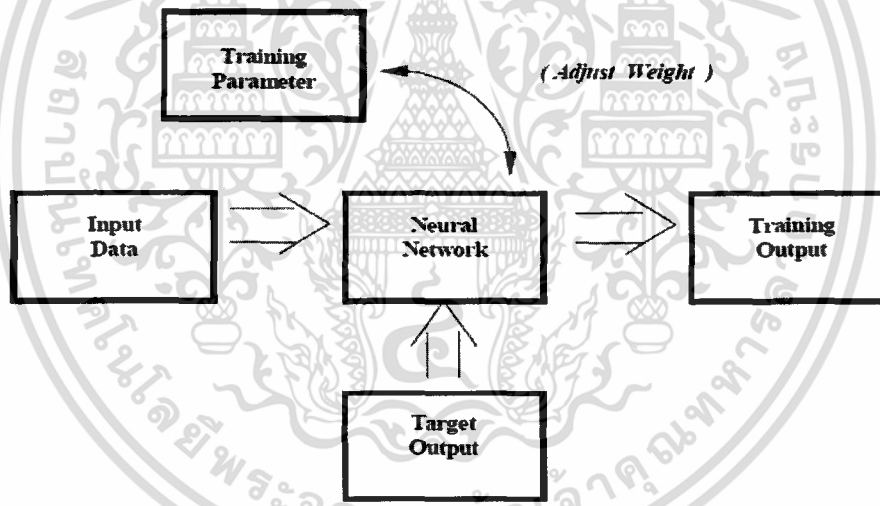
2.2 การจำแนกประเภทข้อมูล

การจำแนกข้อมูล คือการจำแนกข้อมูลแบบรู้ล่วงหน้าว่าข้อมูลมี่กี่ประเภทและประเภทอะไรบ้าง หรือเรียกว่าการเรียนรู้แบบมีผู้สอน(Supervised learning) ซึ่งต้องการผู้สอนเพื่อให้ความรู้กับข้อมูลก่อนว่าประเภทข้อมูลที่ต้องการมีกี่ประเภท ข้อมูลมีลักษณะอย่างไร เพื่อนำการเรียนรู้นั้นไปใช้สำหรับการตัดสินใจ ซึ่งการจำแนกประเภทข้อมูลเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ โดยจะนำข้อมูลส่วนหนึ่งมาสอนให้ระบบเรียนรู้ (Training Data) เพื่อจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ ผลลัพธ์ที่ได้จากการเรียนรู้คือ โมเดลจำแนกประเภทข้อมูล (Classifier Model) และจะนำข้อมูลส่วนที่เหลือจากข้อมูลสอนระบบเป็นข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้องและปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา จะนำข้อมูลมาผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลได้ ซึ่งเทคนิคที่ใช้สำหรับการจำแนกข้อมูลมีให้เลือกใช้หลายเทคนิค เช่น

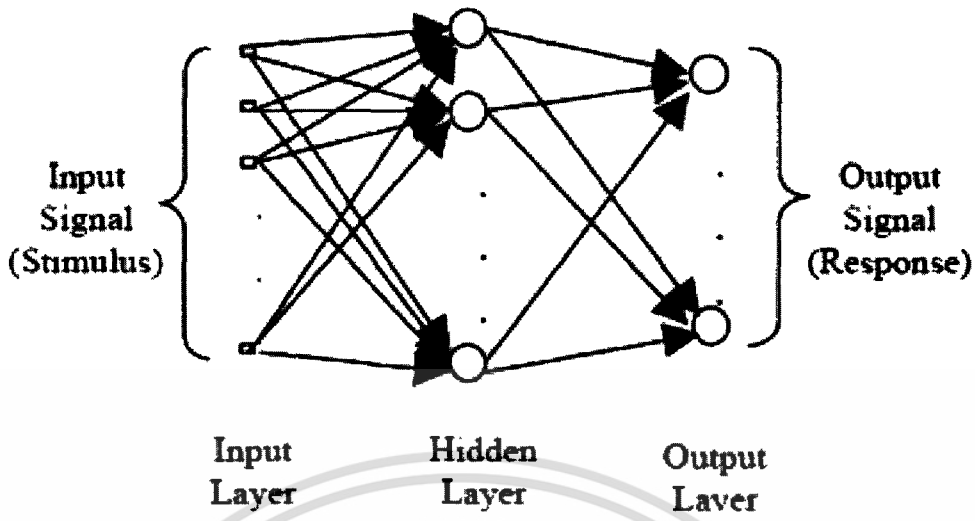
2.2.1 Multilayer Perceptron Neural Network (MLP)

โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น ใช้กับงานที่มีความซับซ้อนได้เป็นอย่างดีประกอบด้วยชั้นต่างๆ คือ ชั้นข้อมูลเข้า(Input Layer) 1 ชั้น ชั้นซ่อน (Hidden Layer) กี่ชั้นก็ได้ และชั้นผลลัพธ์ (Output Layer) 1 ชั้น โดยชั้นซ่อนจะอยู่ระหว่างชั้นข้อมูลเข้าและชั้นผลลัพธ์ ในการเชื่อมต่อระหว่างชั้นต่างๆ ทุกๆ โหนดในชั้นข้อมูลเข้าจะมีเส้นเชื่อมของค่าน้ำหนักเพื่อส่ง

สัญญาณไปยังทุกๆ โหนดในชั้นซ่อน จากนั้นทุกๆ โหนดในชั้นซ่อนจะส่งสัญญาณไปยังทุกๆ โหนดในชั้นผลลัพธ์ โดยมีกระบวนการฝึกฝนเป็นแบบ Supervise ดังรูปที่ 2.3 และใช้ขั้นตอนการส่งค่าย้อนกลับ (Back Propagation) สำหรับการฝึกฝน กระบวนการส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อยคือ การส่งผ่านไปข้างหน้า (Forward Pass) และการส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นของข้อมูลเข้าและจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (error-correction) คือ ผลต่างของผลตอบที่แท้จริง (actual response) กับผลตอบเป้าหมาย (target response) เกิดเป็นสัญญาณผิดพลาด (error signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ ค่าน้ำหนักการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย ดังรูปที่ 2.4



รูปที่ 2.3 แสดงการฝึกฝนแบบมีผู้สอน (Supervise Learning)



รูปที่ 2.4 แสดงโครงข่ายแบบ Multilayer Perceptron Neural Network (MLP) [8]

2.2.2 Naïve Bays

Naïve Bayes เป็นเทคนิคที่ใช้ทฤษฎีของ Bayes Theorem โดยมีความน่าจะเป็นที่จะเกิดเหตุการณ์ต่างๆ ที่ใช้ในการจัดกลุ่มนั้นเป็นอิสระต่อกัน มีการนำวิธีการจำแนกประภทของ Naïve Bays ไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnois) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่น เนื่องจากเป็นวิธีการจำแนกข้อมูลที่มีประสิทธิภาพและมีอัลกอริธึมในการทางานไม่ซับซ้อนเหมือนวิธีการอื่นซึ่งจะวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตาม เพื่อใช้สร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม vj สำหรับข้อมูลที่มีคุณลักษณะทั้งหมด n ตัว ให้ X = {a1, a2,..., an} หรือ ใช้สัญลักษณ์ว่า P(a1, a2,..., an | vj) คือ

$$P(a_1, a_2, \ldots, a_n \mid v_j) = \prod_{i=1}^{n} P(a_i \mid v_j) \qquad (2.6)$$

โดยที่ $\prod$ หมายถึง ผลคูณของค่าทั้งหมด i = 1, 2, 3,..., n และ j = 1, 2, 3,..., n

การนำ Naïve Bays ไปใช้โดยการหาค่าความน่าจะเป็นของค่าที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2, \ldots, a_n \mid v_j)$ จากสมการที่ 2.6 มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้น ๆ คือ $P(V_j)$ ได้เท่ากับ

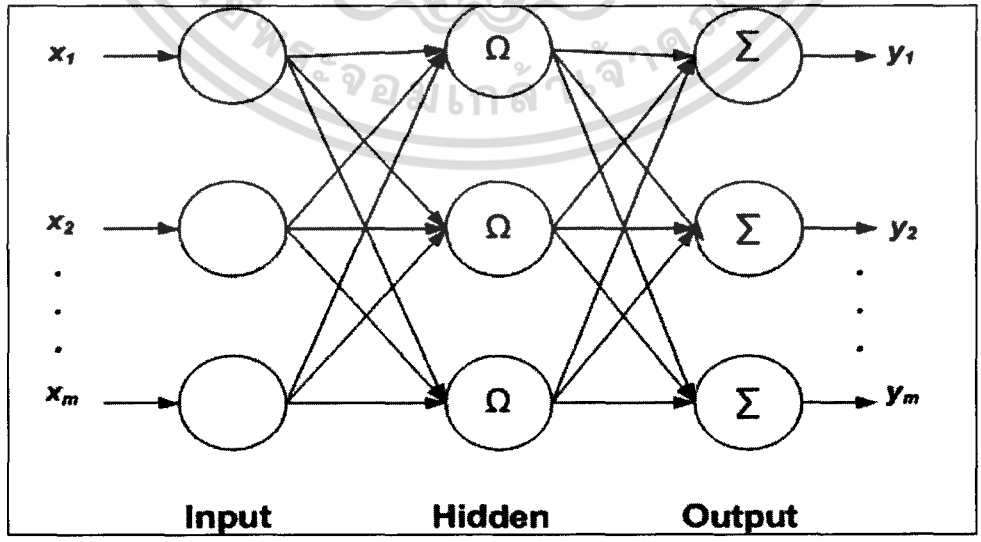
V_{NB} จากนั้นนำค่าที่ได้มาเปรียบเทียบกับกลุ่มที่มีค่าความน่าจะเป็นสูงสุดคือคำตอบ ดังนั้นเราจะได้ว่าวิธีการจำแนกประเภทของ Naïve Bays ดังสมการ

$$v_{NB} = \max P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j) \tag{2.7}$$

2.2.3 Radial Basis Function (RBF)

Radial Basis Function (RBF) เป็นโครงข่ายงานที่มีสถาปัตยกรรมแบบป้อนไปข้างหน้าและมีหลายชั้น จึงมีความสามารถในการทำ non-linear mapping หรือเป็นการแปลงปัญหายากให้เป็นปัญหาที่ง่ายขึ้นและเมื่อลักษณะของปัญหานั้นมีความยากเพิ่มขึ้น การใช้วิธีเพิ่มชั้นของเพอร์เซ็ปตรอนไปเรื่อยๆนั้นอาจทำให้ใช้เวลาค่อนข้างมากในการคำนวณ การลู่เข้าคำตอบก็อาจจะช้าขึ้นและข้อจำกัดของการใช้เส้นแบ่งแยกนั้น ที่สำคัญคือ ไม่อาจจำกัดขอบเขตของกลุ่มของข้อมูลตามธรรมชาติของข้อมูลได้

สถาปัตยกรรมโครงข่าย RBF โดยทั่วไปนั้น เป็นข่ายงานประสาทเทียมแบบป้อนไปข้างหน้าแบบหลายชั้น ซึ่งโครงสร้างของข่ายงานประกอบไปด้วย 3 ชั้น คือ ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นผลลัพธ์ ดังรูปที่ 2.5 ชั้นข้อมูลเข้ามี m นิวรอน ชั้นซ่อนของข่ายงานมี 1 ชั้น สำหรับชั้นผลลัพธ์มี n นิวรอน การเชื่อมต่อระหว่าง ชั้นซ่อนกับ ชั้นผลลัพธ์ จะเชื่อมต่อด้วยค่าน้ำหนัก และมีการปรับค่าน้ำหนักในระหว่างการฝึกสอนข่ายงาน



รูปที่ 2.5 แสดงสถาปัตยกรรมโครงข่ายแบบ RBF [12]

จากรูปที่ 2.5 เมื่อ X_i คือ ข้อมูลเข้า (Input) ที่ถูกส่งค่าไปยังชั้นซ่อน (Hidden Layer) ซึ่งคำนวณได้จาก

$$\theta_j(x) = \frac{x - c_j e}{2\sigma^2 j} \qquad (2.8)$$

เมื่อ θ คือข้อมูลออกที่ j ในชั้นซ่อน, x คือข้อมูลเข้า, C_j คือศูนย์กลางของโหนดที่ j เมื่อกำหนดให้ j =1,2,….n จากนั้นคำนวณค่าข้อมูลออก(Output) ซึ่งคำนวณได้จาก

$$y = \sum_{i=1}^{n} w_i \, \theta_j(x) \qquad (2.9)$$

เมื่อ Wj คือค่าน้ำหนักระหว่างชั้นซ่อนและชั้นข้อมูลออกและ θ_j คือข้อมูลออกจากชั้นซ่อน

บทที่ 3

ขั้นตอนการดำเนินงานวิจัยและการทดลอง

จากบทที่ 2 ได้กล่าวถึงวิธีการคัดเลือกคุณลักษณะที่สำคัญของชุดข้อมูลโดยการใช้กรีดดีอัลกอริธึม ในบทนี้จะกล่าวถึงวิธีการเตรียมข้อมูลก่อนการประมวลผล และวิธีการทดลอง โดยเริ่มจากการเตรียมข้อมูลสำหรับการคัดเลือกคุณลักษณะและการจำแนกประเภทของข้อมูล จากนั้นจะกล่าวถึงวิธีการทดลองทั้งการทดลองการคัดเลือกคุณลักษณะของชุดข้อมูลด้วยวิธีต่างๆ และการนำคุณลักษณะที่ได้ไปจำแนกประเภทข้อมูลด้วยวิธีต่างๆ

3.1 รายละเอียดและที่มาของชุดข้อมูลที่ใช้ในการทดลอง

ในงานวิจัยนี้ได้นำชุดข้อมูลจาก UCI Machine Learning Repository [10] จำนวน 5 ชุด ได้แก่ zoo, pakinsons, hapatitis, network intrusion detection (KDD CUP99) และ thyroid มาใช้ในการทดสอบ ดังตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างชุดข้อมูลสำหรับการทดสอบ

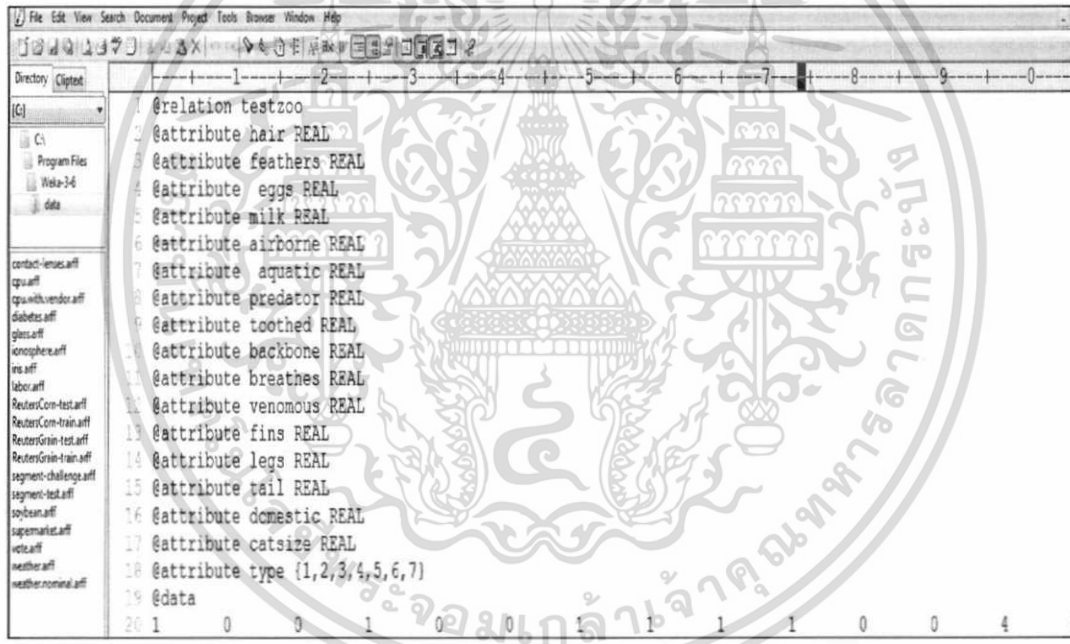
dataset	attributes	instances	attribute type
zoo	17	101	Integer,Nominal
pakinsons	23	197	Real
hapatitis	20	155	Integer,Real,Nominal
network intrusion	42	1000	Integer,Nominal
thyroid	29	9172	Integer,Real,Nominal

จากตารางที่ 3.1 ชุดข้อมูล zoo มีจำนวนคุณลักษณะทั้งหมดเท่ากับ 17 คุณลักษณะ มีรูปแบบการเก็บข้อมูลเป็นสองชนิดคือ Integer และ Nominal และมีข้อมูลจำนวน 101 แถว ชุดข้อมูล pakinsons มีจำนวนคุณลักษณะเท่ากับ 23 คุณลักษณะ มีรูปแบบการเก็บข้อมูลเป็นชนิด Real และมีข้อมูลจำนวน 197 แถว ชุดข้อมูล hepatitis มีจำนวนคุณลักษณะเท่ากับ 20 คุณลักษณะ มีรูปแบบการเก็บข้อมูลเป็นสามชนิดคือ Integer, Real และ Nominal มีข้อมูลจำนวน 155 แถว ชุด

ข้อมูล network intrusion มีจำนวนคุณลักษณะเท่ากับ 42 คุณลักษณะ มีรูปแบบการเก็บข้อมูลเป็นสองชนิดคือ Integer และ Nominal มีข้อมูลจำนวน 4,000,000 แถว แต่งานวิจัยนี้ได้สุ่มเลือกข้อมูลมาจำนวน 1,000 แถวสำหรับเป็นข้อมูลตัวอย่างในการทดสอบ และชุดข้อมูล thyroid มีจำนวนคุณลักษณะเท่ากับ 29 คุณลักษณะ มีรูปแบบการเก็บข้อมูลเป็นสามชนิดคือ Integer, Real และ Nominal มีข้อมูลจำนวน 9,172 แถว

3.2 การเตรียมชุดข้อมูลสำหรับการคัดเลือกคุณลักษณะด้วยโปรแกรม weka 3.6.0

การเตรียมชุดข้อมูลก่อนการประมวลผลด้วยโปรแกรม weka 3.6.0 [11] ต้องกำหนดรูปแบบไฟล์ให้มีนามสกุลเป็น .arff ที่เป็นไฟล์เฉพาะสำหรับโปรแกรม weka ซึ่งงานวิจัยนี้ได้เลือกใช้โปรแกรม Editplus 3 สำหรับขั้นตอนการเตรียมข้อมูลทดสอบ ดังรูปที่ 3.1



รูปที่ 3.1 แสดงการเตรียมข้อมูลด้วยโปรแกรม Editplus 3

จากรูปที่ 3.1 บรรทัดแรกหมายถึงการกำหนดชื่อตารางให้กับชุดข้อมูลโดยกำหนดให้ขึ้นต้นด้วย @ relation ในตัวอย่างนี้คือชุดข้อมูล zoo มีการกำหนดชื่อตารางเป็น testzoo และบรรทัดที่ 2 – 18 คือการกำหนดชื่อคุณลักษณะของข้อมูลโดยกำหนดให้ขึ้นต้นด้วย @ attribute ตามด้วยชื่อคุณลักษณะและชนิดข้อมูลของแต่ละคุณลักษณะ ส่วนบรรทัดที่ 19 คือการบอกให้โปรแกรมรู้ว่าตั้งแต่บรรทัดนี้เป็นต้นไปจะเป็นข้อมูลที่จะใช้สำหรับการทดสอบโดยกำหนดให้ขึ้นต้นด้วย @ data และตามด้วยข้อมูลแถวแรก ลักษณะการเก็บข้อมูลดังตัวอย่างต่อไปนี้

1 0 0 1 0 0 1 1 1 1 0 0 4 0 0 1 1 ข้อมูลแถวที่หนึ่ง

1 0 0 1 0 0 0 1 1 1 0 0 4 1 0 1 1 ข้อมูลแถวที่สอง

จากตัวอย่างข้อมูลข้างต้น ข้อมูลตัวแรก หมายถึงคุณลักษณะตัวที่หนึ่ง ข้อมูลตัวที่สอง หมายถึงคุณลักษณะตัวที่สอง ตามลำดับไปเรื่อยๆ จนครบจำนวน 17 คุณลักษณะของชุดข้อมูลนี้ (ข้อมูลหนึ่งบรรทัดหมายถึงข้อมูลหนึ่งแถวในฐานข้อมูล) เมื่อใส่ข้อมูลครบแล้วให้บันทึกไฟล์งานที่ได้เป็นไฟล์ .arff จากนั้นก็นำไฟล์ที่ได้ไปประมวลผลด้วยโปรแกรม weka

ขั้นตอนการประมวลผลเริ่มตั้งแต่การคัดเลือกคุณลักษณะของชุดข้อมูล โดยงานวิจัยนี้ได้เลือกใช้วิธีการคัดเลือกคุณลักษณะด้วยกรีดดีอัลกอริธึม สามารถเรียกใช้ได้จากเมนู Select Attribute/GreedyStepwise และเมื่อคัดเลือกคุณลักษณะของชุดข้อมูลด้วยกรีดดีอัลกอริธึมแล้ว จะทำให้ได้เฉพาะคุณลักษณะที่สำคัญที่สุด ดังรูปที่ 3.2

```
Selected attributes: 1,2,4,8,9,10,12,13,14 : 9
                     hair
                     feathers
                     milk
                     toothed
                     backbone
                     breathes
                     fins
                     legs
                     tail
```

รูปที่ 3.2 แสดงจำนวนคุณลักษณะที่ได้จากการคัดเลือกด้วยกรีดดีอัลกอริธึม

จากรูปที่ 3.2 การคัดเลือกคุณลักษณะโดยใช้กรีดดีอัลกอริธึมจะทำให้ได้เฉพาะคุณลักษณะที่สำคัญเท่านั้น และสามารถนำไปใช้งานต่อได้ทันที ซึ่งจะต่างกับวิธีการคัดเลือกคุณลักษณะของ Information Gain และ Principal Component Analysis ที่ให้ค่าของคุณลักษณะแต่ละตัวเรียงตามลำดับค่าความสำคัญของคุณลักษณะแต่ละตัว ดังรูปที่ 3.3 และ รูปที่ 3.4

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 type):
        Information Gain Ranking Filter

Ranked attributes:
 1.311   13 legs
 0.974    4 milk
 0.866    8 toothed
 0.83     3 eggs
 0.791    1 hair
 0.718    2 feathers
 0.676    9 backbone
 0.614   10 breathes
 0.5     14 tail
 0.47     5 airborne
 0.467   12 fins
 0.389    6 aquatic
 0.308   16 catsize
 0       11 venomous
 0        7 predator
 0       15 domestic

Selected attributes: 13,4,8,3,1,2,9,10,14,5,12,6,16,11,7,15 : 16
```

รูปที่ 3.3 แสดงการคัดเลือกคุณลักษณะด้วย Information Gain

จากรูปที่ 3.3 การคัดเลือกคุณลักษณะโดยใช้ Information Gain จะทำให้ได้ค่าความสำคัญของคุณลักษณะแต่ละตัว โดยเรียงลำดับจากคุณลักษณะที่มีค่าความสำคัญสูงสุดเรียงลำดับไปเรื่อยๆ จนถึงคุณลักษณะที่มีค่าความสำคัญน้อยที่สุด (สามารถดูรายละเอียดการกำหนดเกณฑ์การเลือกคุณลักษณะได้ในบทที่ 4)

```
Ranked attributes:
 0.7081    1 0.445milk-0.432eggs+0.408hair+0.323toothed+0.289catsize...
 0.4993    2 0.455fins+0.374aquatic-0.366breathes-0.341legs+0.321toothed...
 0.3515    3 0.483tail+0.483feathers+0.476backbone-0.313legs+0.278airborne...
 0.2745    4 -0.628domestic+0.578predator+0.298catsize-0.216fins+0.178feathers...
 0.2149    5 -0.899venomous-0.227predator-0.194tail-0.133backbone+0.13 fins...
 0.1683    6 0.703domestic+0.457catsize+0.346predator-0.199toothed+0.196feathers...
 0.1331    7 0.688catsize-0.468predator+0.3  venomous+0.253fins-0.233domestic...
 0.1011    8 0.67 airborne+0.48 aquatic+0.335hair+0.26 fins+0.199milk...
 0.0731    9 0.659legs+0.615tail-0.277breathes-0.215predator+0.209aquatic...
 0.0494   10 0.547aquatic+0.405breathes+0.357backbone-0.317predator-0.304tail...

Selected attributes: 1,2,3,4,5,6,7,8,9,10 : 10
```

รูปที่ 3.4 แสดงการคัดเลือกคุณลักษณะด้วย Principal Component Analysis

จากรูปที่ 3.4 การคัดเลือกคุณลักษณะโดยใช้ Principal Component Analysis สามารถกำหนดจำนวนคุณลักษณะที่ต้องการได้ ซึ่งตัวอย่างในรูปนี้ได้มีการกำหนดจำนวนคุณลักษณะเท่ากับ 5คุณลักษณะ (สามารถดูรายละเอียดการกำหนดเกณฑ์การกำหนดจำนวนคุณลักษณะได้ในบทที่ 4)

เมื่อคัดเลือกคุณลักษณะของชุดข้อมูลด้วยกรีดดีอัลกอริธึมครบทุกชุดแล้ว จำนวนคุณลักษณะที่ได้จากการคัดเลือกคุณลักษณะ ดังตารางที่ 3.2

ตารางที่ 3.2 แสดงจำนวนคุณลักษณะที่ได้จากการคัดเลือกด้วยกรีดดีอัลกอริธึม

dataset	Before Select	After Select
zoo	17	9
pakinsons	23	10
hapatitis	20	10
network intrusion	42	6
thyroid	29	11

จากตารางที่ 3.2 การคัดเลือกคุณลักษณะด้วยกรีดดีอัลกอริธึมสามารถลดจำนวนคุณลักษณะลงได้เท่ากับ 9, 10, 10, 6, 11 ตามลำดับ จากนั้นนำคุณลักษณะของชุดข้อมูลที่ได้จากการคัดเลือกคุณลักษณะดังกล่าวไปจำแนกประเภทข้อมูลต่อไป

3.3 การเตรียมข้อมูลสำหรับการจำแนกข้อมูล

จากหัวข้อที่ 3.2 เมื่อคัดเลือกคุณลักษณะด้วยวิธีต่างๆ จะทำให้ได้จำนวนคุณลักษณะที่ต่างกันจากนั้นให้ทำการเลือกคุณลักษณะที่ไม่สำคัญออกจากจำนวนคุณลักษณะทั้งหมดของชุดข้อมูลนั้นๆ จากนั้นก็นำคุณลักษณะที่ได้ไปจำแนกข้อมูลด้วย Radial Basis Function, Multi – Layer Perceptron และ Naïve Bays โดยให้เลือกที่เมนู Classify/Function/Radial Basis Function

ชุดข้อมูลอื่นๆ ที่ใช้ในการทดสอบครั้งนี้ ผู้วิจัยได้ทำตามขั้นตอนในหัวข้อที่ 3.2 และ 3.3 จนครบทุกชุดข้อมูล การคัดเลือกคุณลักษณะมีการทดลองกับทั้งสามอัลกอริธึมคือ Greedy Algorithms, Information Gain และ Principal Component Analysis ส่วนการจำแนกประเภทข้อมูล

ก็มีการทดลองกับทั้งสามอัลกอริธึมเช่นกันคือ Radial Basis Function, Multi- Layer Perceptron และ Naive Bays

3.4 วิธีการทดลอง

ในการทดลองได้แบ่งการทดลอง ดังนี้

1. ทดลองเพื่อเปรียบเทียบประสิทธิภาพและเวลาในจำแนกข้อมูลโดยไม่มีการคัดเลือกคุณลักษณะของชุดข้อมูลโดยใช้ Radial Basis Function, Multi- Layer Perceptron และ Naive Bays ซึ่งการทดลองในขั้นตอนนี้จะใช้จำนวนคุณลักษณะทั้งหมดของชุดข้อมูลในการทดลอง โดยไม่มีการคัดเลือกคุณลักษณะ เพื่อนำค่าความถูกต้องที่ได้ไปเปรียบเทียบกับการจำแนกประเภทข้อมูลที่มีการคัดเลือกคุณลักษณะต่อไป สามารถดูผลการเปรียบเทียบได้ในบทที่ 4

2. ทดลองเพื่อเปรียบเทียบประสิทธิภาพในการคัดเลือกคุณลักษณะด้วย Greedy Algorithms, Information Gain และ Principal Component Analysis ก่อนการจำแนกประเภทข้อมูลด้วย Radial Basis Function การทดลองในขั้นตอนนี้จะใช้อัลกอริธึมสำหรับการคัดเลือกคุณลักษณะสามอัลกอริธึม ซึ่งแต่ละอัลกอริธึมจะได้คุณลักษณะที่ต่างกัน ดังตารางที่ 3.3 – 3.5

ตารางที่ 3.3 แสดงคุณลักษณะที่ได้จากการคัดเลือกโดยใช้กรีดดีอัลกอริธึม

Data Set	Attributes
zoo	hair, feathers, milk, toothed, backbone, breathes, fins, legs, tail
pakinsons	Fo, Fhi, Flo, RAP, MDVP_APQ, NHR, spread1, spread2, D2, PPE
hapatitis	AGE, SEX, MALAISE, SPIDERS, ASCITES, VARICES, BILIRUBIN, ALBUMIN, PROTIME, HISTOLOGY
Network intrusion	service, src_bytes, srv_count, same_srv_rate, diff_srv_rate
thyroid	age,sex, on_thyroxine, sick, I131_treatment, query_hyperthyroid, lithium, psych, TSH_measured, T4U

จากตารางที่ 3.3 การคัดเลือกคุณลักษณะโดยใช้กรีดดีอัลกอริธึมทำให้ได้เฉพาะคุณลักษณะที่สำคัญที่สุด และสามารถนำคุณลักษณะนั้นไปใช้ในการจำแนกข้อมูลได้ทันที ซึ่งจะต่างกับวิธีการคัดเลือกคุณลักษณะโดยใช้ Information Gain และ Principal Component Analysis

ตารางที่ 3.4 แสดงจำนวนคุณลักษณะที่ได้จากการคัดเลือกโดยใช้ Information Gain

Data Set	Attributes
zoo	legs, milk, toothed, eggs, hair, feathers, backbone, breathes
pakinsons	PPE, spread1, Fo, Jitter2, MDVP_APQ, Shimmer_APQ5, MDVP_Shimmer1, Flo
hapatitis	ALBUMIN, BILIRUBIN, ASCITES, SPIDERS, HISTOLOGY, FATIGUE, MALAISE
Network	service, src_bytes, dst_host_srv_count, dst_host_diff_srv_rate, count
thyroid	psych,age,T3,T4U,on_thyroxine,T3_measured,TSH_measured,sex,TT4,TT4_measured, TSH,pregnant,lithium,FTI_measured,T4U_measured,sick,TBG_measured, FTI

จากตารางที่ 3.4 การคัดเลือกคุณลักษณะโดยใช้ Information Gain ซึ่งข้อมูลแต่ละชุดจะใช้เกณฑ์ในการเลือกที่ต่างกัน โดยชุดข้อมูล zoo ใช้เกณฑ์การเลือกจากค่าความสำคัญของคุณลักษณะที่มีค่าตั้งแต่ 0.61 ชุดข้อมูล pakinsons ใช้เกณฑ์การเลือกจากค่าความสำคัญของคุณลักษณะที่มีค่าตั้งแต่ 0.18 ชุดข้อมูล hepatitis ใช้เกณฑ์การเลือกจากค่าความสำคัญของคุณลักษณะที่มีค่าตั้งแต่ 0.08 ชุดข้อมูล network intrusion ใช้เกณฑ์การเลือกจากค่าความสำคัญของคุณลักษณะที่มีค่าตั้งแต่ 0.93 และชุดข้อมูล thyroid ใช้เกณฑ์การเลือกจากค่าความสำคัญของคุณลักษณะที่มีค่าตั้งแต่ 0.03 ซึ่งการเลือกคุณลักษณะด้วยวิธีนี้อาจจะไม่ใช่วิธีที่เหมาะสมจึงอาจทำให้เกิดความผิดพลาดด้านประสิทธิภาพได้ ซึ่งจะต่างจาก Greedy Algorithms ที่ได้เฉพาะคุณลักษณะที่สำคัญเท่านั้น และ Principal Component Analysis ที่สามารถกำหนดจำนวนคุณลักษณะที่ต้องการได้

ตารางที่ 3.5 แสดงจำนวนคุณลักษณะที่ได้จากการคัดเลือกโดยใช้ Principal Component Analysis

Data Set	Attributes
zoo	milk, eggs, hair, toothed, catsize, breathes, backbone, aquatic, feathers
pakinsons	MDVP_Shimmer2, MDVP_Shimmer1, PPQ, Jitter1, MDVP_APQ, Shimmer_DDA, Shimmer_APQ3, Shimmer_APQ5, RAP
hapatitis	ALBUMIN, BILIRUBIN, SPIDERS, FATIGUE
Network intrusion	rerror_rate, diff_srv_rate, SF, dst_host_diff_srv_rate, dst_host_rerror_rate, same_srv_rate,srv_rerror_rate, flag,dst_host_srv_rerror_rate, dst_host_same_srv_rate
thyroid	TT4_measured, FTI_measured, T4U_measured, TBG_measured, TSH_measured, T3_measured, tumor, sex, psych, query_hyperthyroid, lithium, on_thyroxine

จากตารางที่ 3.5 การคัดเลือกคุณลักษณะด้วย Principal Component Analysis สามารถกำหนดจำนวนคุณลักษณะที่ต้องการได้ ซึ่งงานวิจัยนี้ได้กำหนดเกณฑ์การเลือกดังนี้ ชุดข้อมูล zoo ใช้เกณฑ์การเลือกจากจำนวนคุณลักษณะที่ 9 คุณลักษณะ ชุดข้อมูล pakinsons ใช้เกณฑ์การเลือกจากจำนวนคุณลักษณะที่ 9 คุณลักษณะ ชุดข้อมูล hepatitis ใช้เกณฑ์การเลือกจากจำนวนคุณลักษณะที่ 8 คุณลักษณะ ชุดข้อมูล network intrusion ใช้เกณฑ์การเลือกจากจำนวนคุณลักษณะที่ 10 คุณลักษณะ และชุดข้อมูล thyroid ใช้เกณฑ์การเลือกจากจำนวนคุณลักษณะที่ 12 คุณลักษณะ

ประสิทธิภาพของการคัดเลือกคุณลักษณะของชุดข้อมูลแต่ละอัลกอริธึม สามารถดูผลการเปรียบเทียบได้ในบทที่ 4

3. ทดลองเพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลด้วย Radial Basis Function, Multi- Layer Perceptron และ Naive Bays โดยใช้คุณลักษณะทั้งหมดของชุดข้อมูล หลังจากนั้นทดลองการจำแนกประเภทข้อมูลโดยใช้การคัดเลือกคุณลักษณะด้วย Greedy Algorithms แล้วนำประสิทธิภาพของทั้งสองวิธีมาเปรียบเทียบกัน ซึ่งสามารถดูผลการเปรียบเทียบได้ในบทที่ 4

3.5 การวัดประสิทธิภาพ

งานวิจัยนี้ใช้การทดลองแบบ K-fold cross –validation เพื่อแบ่งชุดข้อมูลสำหรับการสอนและการทดสอบ ดังนั้นแต่ละรอบของการทดลองจะมีการคำนวณหาค่าความแม่นยำเสมอ ดังสมการ

$$\text{accuracy}_i = \frac{\text{Correctly Classified Instances}}{\text{Total Number of Instances}} \times 100 \qquad (3.1)$$

หลังจากทำการทดลองครบทุกรอบ จะต้องหาค่าเฉลี่ยของความแม่นยำเพื่อเป็นผลลัพธ์สุดท้าย ซึ่งจะนำมาใช้สำหรับเปรียบเทียบประสิทธิภาพ

บทที่ 4

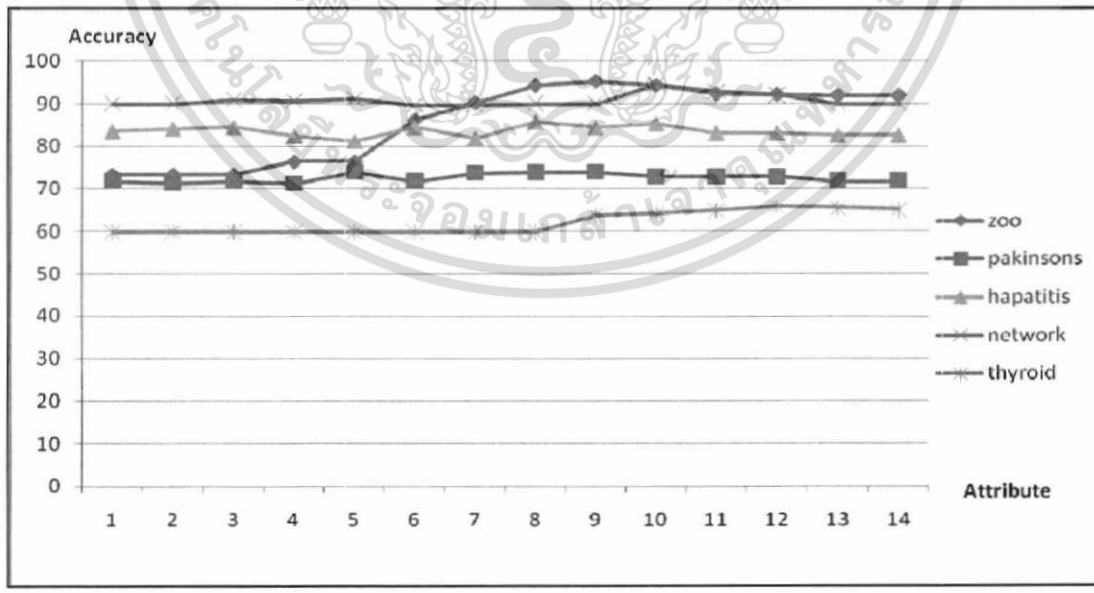
ผลการทดลอง

ในบทที่ 3 ได้กล่าวถึงวิธีและขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล ในบทนี้จะแสดงผลการทดลองโดยแบ่งเป็นสามส่วนคือ ส่วนแรกจะแสดงผลการทดลองการกำหนดเกณฑ์การคัดเลือกคุณลักษณะด้วย Principal Component Analysis และ Information Gain ส่วนที่สองจะแสดงผลการทดลองที่ได้จากการเปรียบเทียบประสิทธิภาพและเวลาที่ใช้ในการประมวลผลสำหรับการจำแนกประเภทข้อมูลโดยการใช้คุณลักษณะทั้งหมดของชุดข้อมูล ในส่วนที่สามจะแสดงผลการทดลองที่ได้จากการเปรียบเทียบประสิทธิภาพและเวลาที่ใช้ในการประมวลผลสำหรับการจำแนกข้อมูลที่มีการคัดเลือกคุณลักษณะก่อนการจำแนกประเภทข้อมูล

4.1 ผลการทดลองการกำหนดเกณฑ์การคัดเลือกคุณลักษณะ

4.1.1 การกำหนดเกณฑ์การคัดเลือกคุณลักษณะด้วย Principal Component Analysis

ผลจากการทดลองการกำหนดเกณฑ์การคัดเลือกคุณลักษณะด้วยเทคนิคของ Principal Component Analysis ที่สามารถกำหนดจำนวนคุณลักษณะที่ต้องการได้ ผลการทดลองดังรูปที่ 4.1

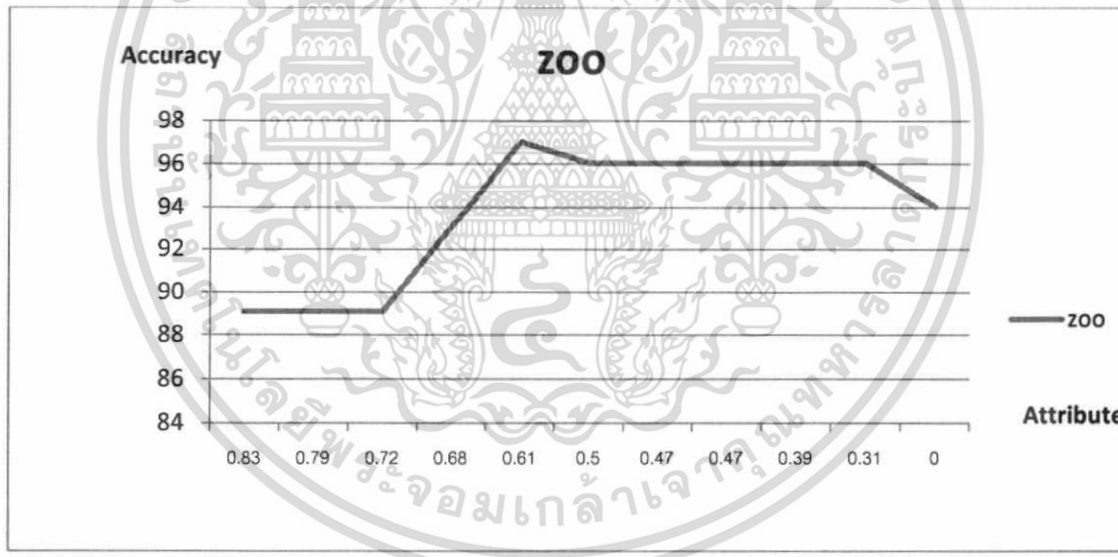


รูปที่ 4.1 แสดงผลการกำหนดเกณฑ์การคัดเลือกคุณลักษณะด้วย Principal Component Analysis

จากรูปที่ 4.1 การคัดเลือกคุณลักษณะของชุดข้อมูลแต่ละชุด มีการกำหนดเกณฑ์การคัดเลือกคุณลักษณะที่ต่างกันคือ ชุดข้อมูล zoo มีค่าความถูกต้องของข้อมูลสูงสุดที่คุณลักษณะจำนวน 9 คุณลักษณะ ชุดข้อมูล pakinsons มีค่าความถูกต้องของข้อมูลสูงสุดที่คุณลักษณะจำนวน 9 คุณลักษณะ ชุดข้อมูล hapatitis มีค่าความถูกต้องของข้อมูลสูงสุดที่คุณลักษณะจำนวน 8 คุณลักษณะ ชุดข้อมูล network intrusion มีค่าความถูกต้องของข้อมูลสูงสุดที่คุณลักษณะจำนวน 10 คุณลักษณะ ชุดข้อมูล thyroid มีค่าความถูกต้องของข้อมูลสูงสุดที่คุณลักษณะจำนวน 15 คุณลักษณะ

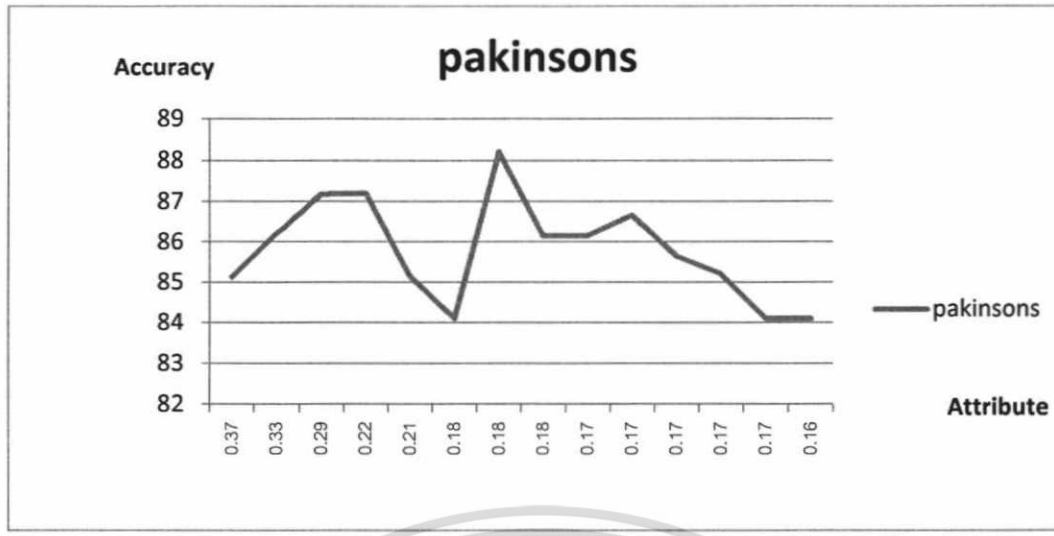
4.1.2 การกำหนดเกณฑ์การคัดเลือกคุณลักษณะด้วย Information Gain

ผลจากการทดลองการกำหนดเกณฑ์การคัดเลือกคุณลักษณะด้วยเทคนิคของ Information Gain ที่ให้ผลการคัดเลือกคุณลักษณะคือ การเรียงลำดับตามค่าความสำคัญของคุณลักษณะแต่ละตัว ผลการทดลองดังรูปที่ 4.2 – 4.6



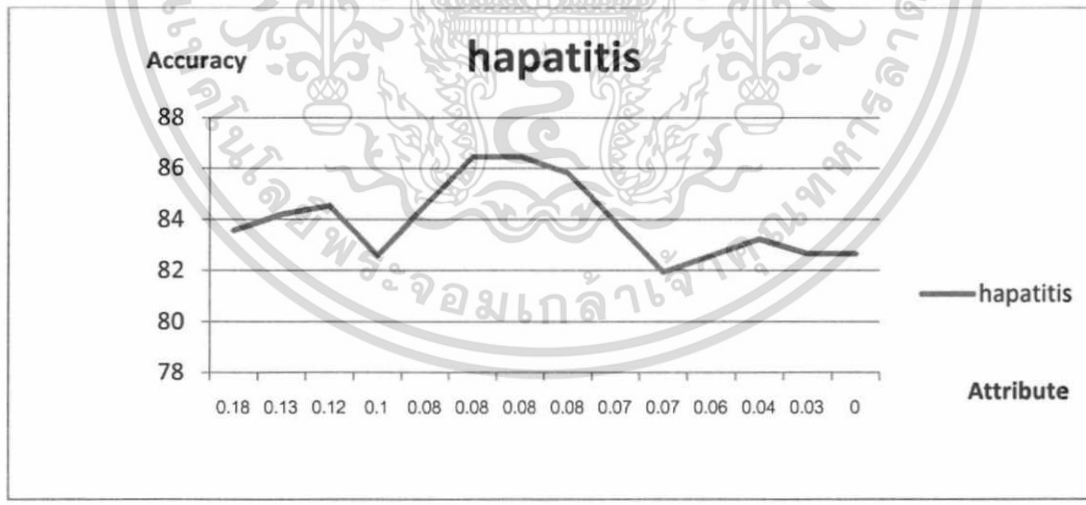
รูปที่ 4.2 แสดงผลการกำหนดเกณฑ์การคัดเลือกคุณลักษณะของชุดข้อมูล zoo ด้วย Information Gain

จากรูปที่ 4.2 การกำหนดเกณฑ์การเลือกคุณลักษณะของชุดข้อมูล zoo ที่ได้จากการคัดเลือกด้วย Information Gain จะเห็นว่าการกำหนดเกณฑ์การเลือกที่ค่าความสำคัญเท่ากับ 0.61 จะให้ค่าความถูกต้องของข้อมูลสูงที่สุด และเมื่อค่าความสำคัญของคุณลักษณะต่ำลงค่าความถูกต้องของข้อมูลก็จะลดลงเรื่อยๆ



รูปที่ 4.3 แสดงผลการกำหนดเกณฑ์การคัดเลือกคุณลักษณะของชุดข้อมูล pakinsons ด้วย Information Gain

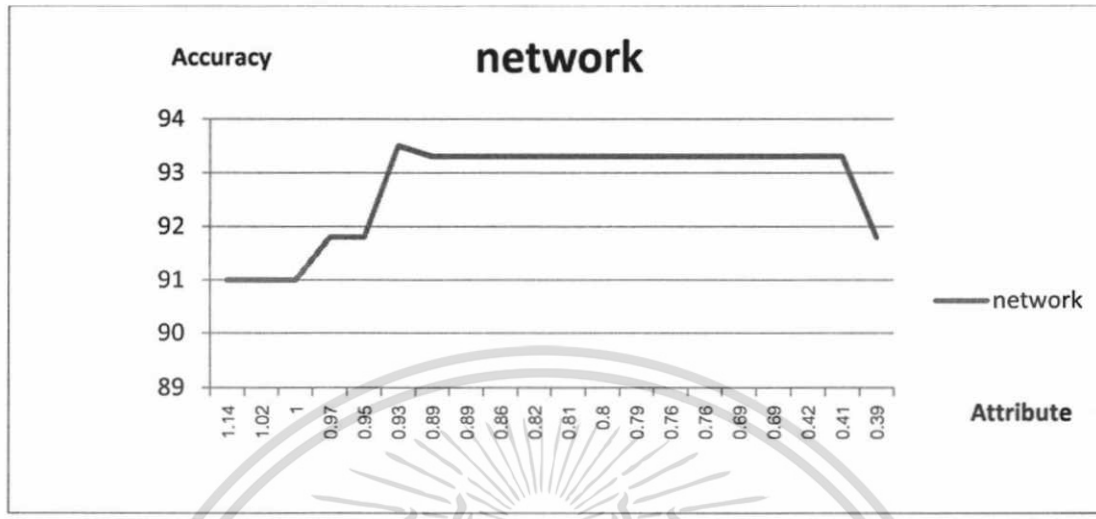
จากรูปที่ 4.3 การกำหนดเกณฑ์การเลือกคุณลักษณะของชุดข้อมูล pakinsons ที่ได้จากการคัดเลือกด้วย Information Gain จะเห็นว่าการกำหนดเกณฑ์การเลือกที่ค่าความสำคัญเท่ากับ 0.18 จะให้ค่าความถูกต้องของข้อมูลสูงที่สุด และเมื่อค่าความสำคัญของคุณลักษณะต่ำลงค่าความถูกต้องของข้อมูลก็จะลดลงเรื่อยๆ



รูปที่ 4.4 แสดงผลการกำหนดเกณฑ์การคัดเลือกคุณลักษณะของชุดข้อมูล hepatitis ด้วย Information Gain

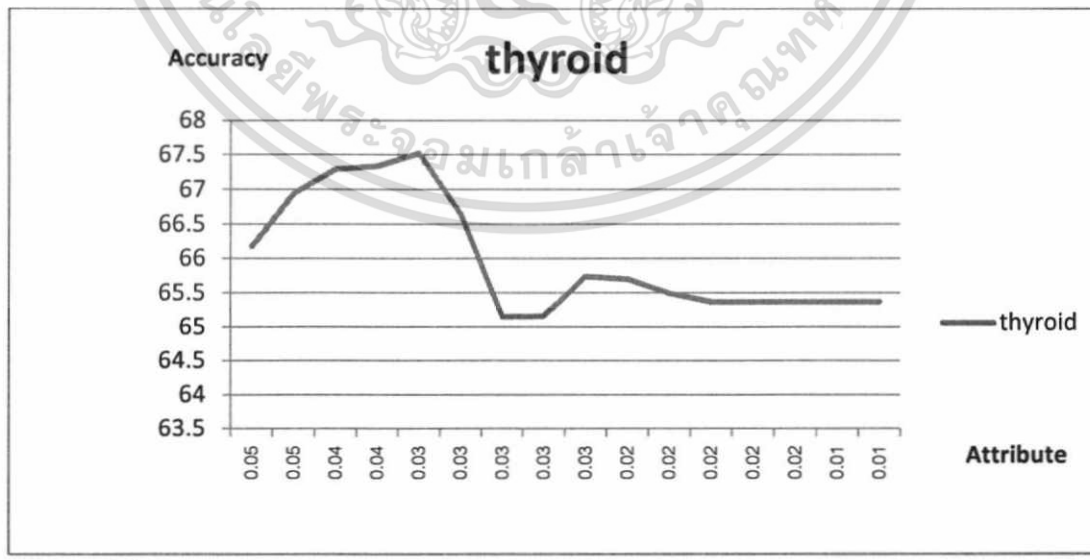
จากรูปที่ 4.4 การกำหนดเกณฑ์การเลือกคุณลักษณะของชุดข้อมูล hapatitis ที่ได้จากการคัดเลือกด้วย Information Gain จะเห็นว่าการกำหนดเกณฑ์การเลือกที่ค่าความสำคัญเท่ากับ 0.08 จะ

ให้ค่าความถูกต้องของข้อมูลสูงที่สุด และเมื่อค่าความสำคัญของคุณลักษณะต่ำลงค่าความถูกต้องของข้อมูลก็จะลดลงเรื่อยๆ



รูปที่ 4.5 แสดงผลการกำหนดเกณฑ์การคัดเลือกคุณลักษณะของชุดข้อมูล network intrusion ด้วย Information Gain

จากรูปที่ 4.5 การกำหนดเกณฑ์การเลือกคุณลักษณะของชุดข้อมูล network intrusion ที่ได้จากการคัดเลือกด้วย Information Gain จะเห็นว่าการกำหนดเกณฑ์การเลือกที่ค่าความสำคัญเท่ากับ 0.93 จะให้ค่าความถูกต้องของข้อมูลสูงที่สุด และเมื่อค่าความสำคัญของคุณลักษณะต่ำลงค่าความถูกต้องของข้อมูลก็จะลดลงเรื่อยๆ

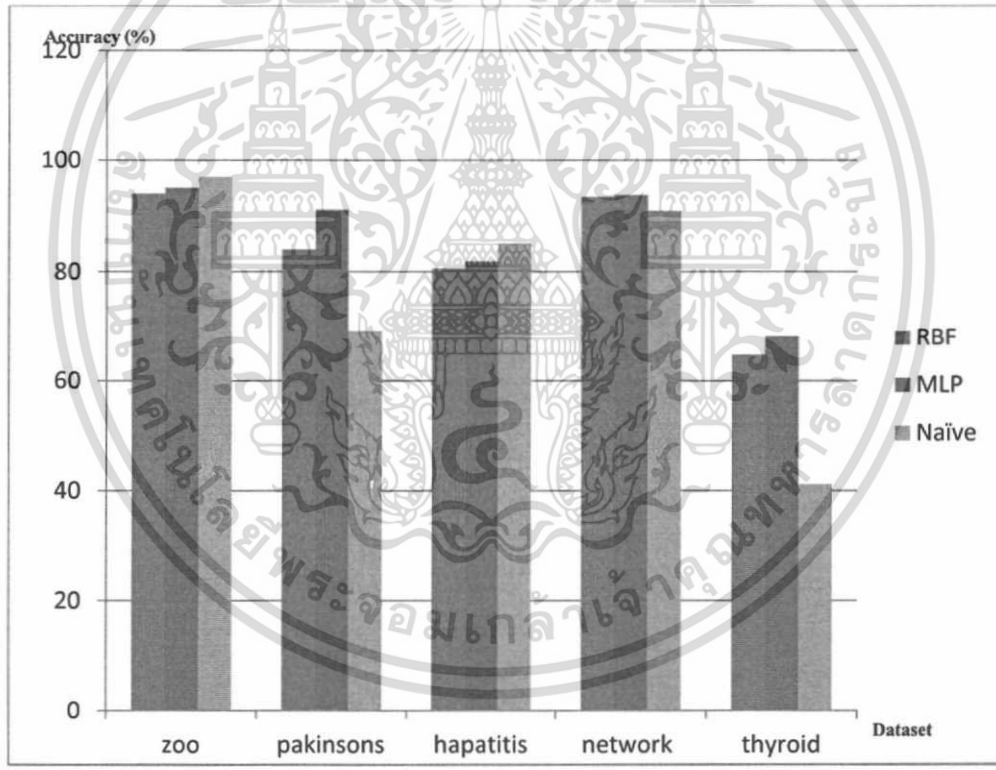


รูปที่ 4.6 แสดงผลการกำหนดเกณฑ์การคัดเลือกคุณลักษณะของชุดข้อมูล thyroid ด้วย Information Gain

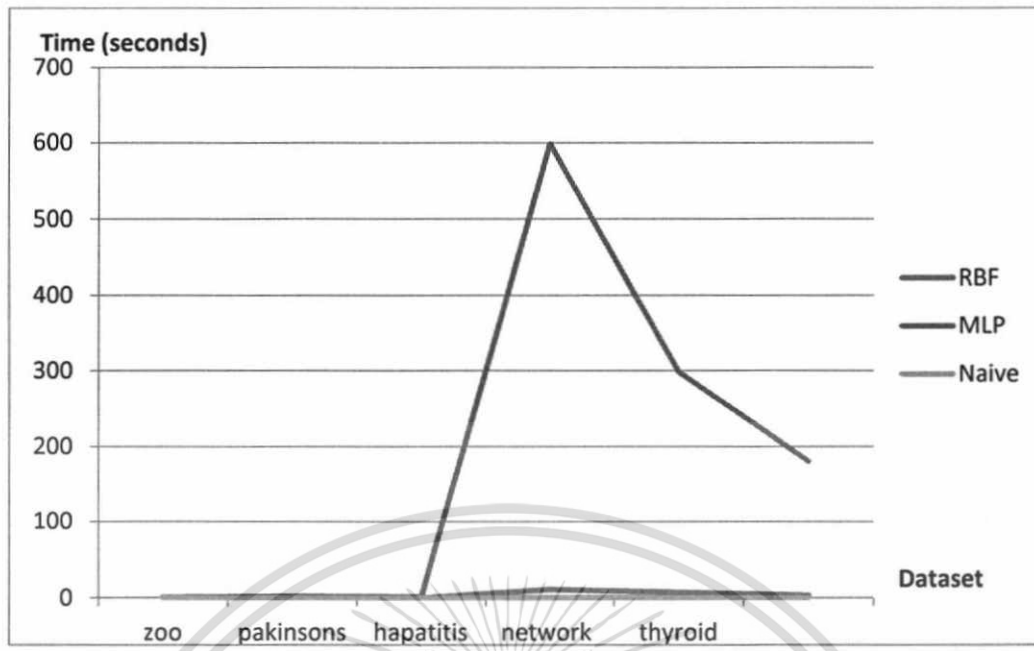
จากรูปที่ 4.6 การกำหนดเกณฑ์การเลือกคุณลักษณะของชุดข้อมูล thyroid ที่ได้จากการคัดเลือกด้วย Information Gain จะเห็นว่าการกำหนดเกณฑ์การเลือกที่ค่าความสำคัญเท่ากับ 0.03 จะให้ค่าความถูกต้องของข้อมูลสูงที่สุด และเมื่อค่าความสำคัญของคุณลักษณะต่ำลงค่าความถูกต้องของข้อมูลก็จะลดลงเรื่อยๆ

4.2 ผลการทดลองการจำแนกข้อมูลโดยการใช้คุณลักษณะทั้งหมด

ผลจากการทดลองการจำแนกข้อมูลที่มีการใช้คุณลักษณะทั้งหมดของชุดข้อมูล โดยใช้เทคนิคการจำแนกข้อมูลด้วย Radial Basis Function, Multilayer Perceptron และ Naïve Bays เมื่อเปรียบเทียบค่าความถูกต้องและเวลาที่ใช้ในการประมวลผลที่ได้จากการจำแนกข้อมูลโดยใช้คุณลักษณะทั้งหมดด้วยอัลกอริธึมทั้งสามในรูปแบบของกราฟได้ดังรูปที่ 4.7 และรูปที่ 4.8



รูปที่ 4.7 เปรียบเทียบค่าความถูกต้องที่ได้จากการจำแนกข้อมูลด้วยสามอัลกอริธึม

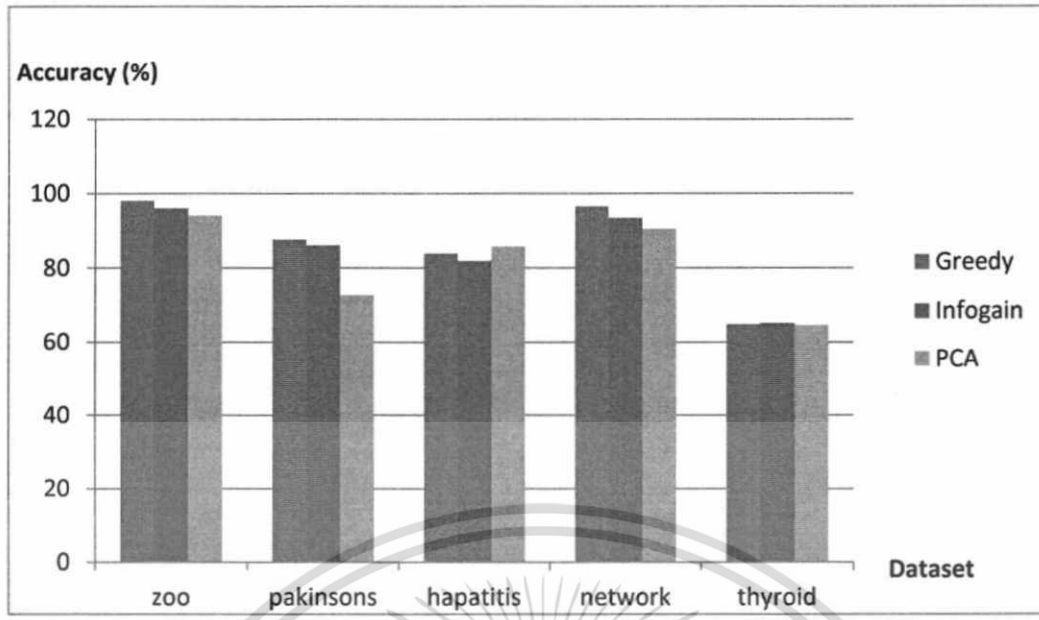


รูปที่ 4.8 เปรียบเทียบเวลาที่ใช้ในการประมวลผลด้วยสามอัลกอริธึม

จากรูปที่ 4.7 และรูปที่ 4.8 เมื่อพิจารณาทั้งสองด้านร่วมกันจะเห็นว่า เทคนิคของ Naïve Bays ให้ค่าความถูกต้องที่ค่อนข้างต่ำแต่ใช้เวลาในการประมวลผลน้อย ส่วนเทคนิคของ Multilayer Perceptron จะให้ค่าความถูกต้องสูงแต่ใช้เวลาในการประมวลผลมาก ส่วนเทคนิคของ Radial Basis Function ที่ให้ค่าความถูกต้องในระดับที่ค่อนข้างสูงและใช้เวลาในการประมวลผลน้อย จะสามารถนำมาแก้ปัญหาทั้งสองด้านนี้ได้ ดังนั้นงานวิจัยนี้จึงเลือกใช้เทคนิคการจำแนกข้อมูลด้วย Radial Basis Function ในการจำแนกข้อมูลหลังจากการคัดเลือกคุณลักษณะต่อไป

4.3 เปรียบเทียบประสิทธิภาพการคัดเลือกคุณลักษณะ

ในหัวข้อนี้เป็นการการเปรียบเทียบประสิทธิภาพการคัดเลือกคุณลักษณะของข้อมูล โดยใช้เทคนิคการคัดเลือกคุณลักษณะด้วย Greedy Algorithms, Information Gain และ Principal Component Analysis จากนั้นนำคุณลักษณะที่ได้จากทั้งสามอัลกอริธึม ไปจำแนกข้อมูลด้วยเทคนิค Radial Basis Function ที่ได้เลือกไว้ในหัวข้อที่ 4.2 ผลการทดลองดังรูปที่ 4.9

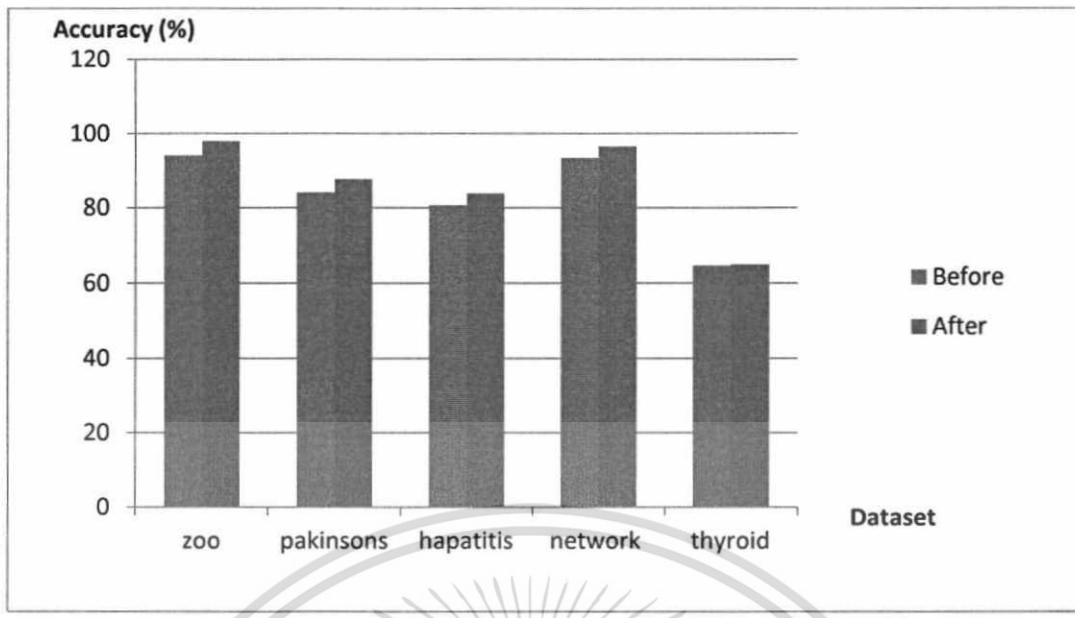


รูปที่ 4.9 แสดงการคัดเลือกคุณลักษณะด้วย Greedy Algorithms, Information Gain และ Principal Component Analysis

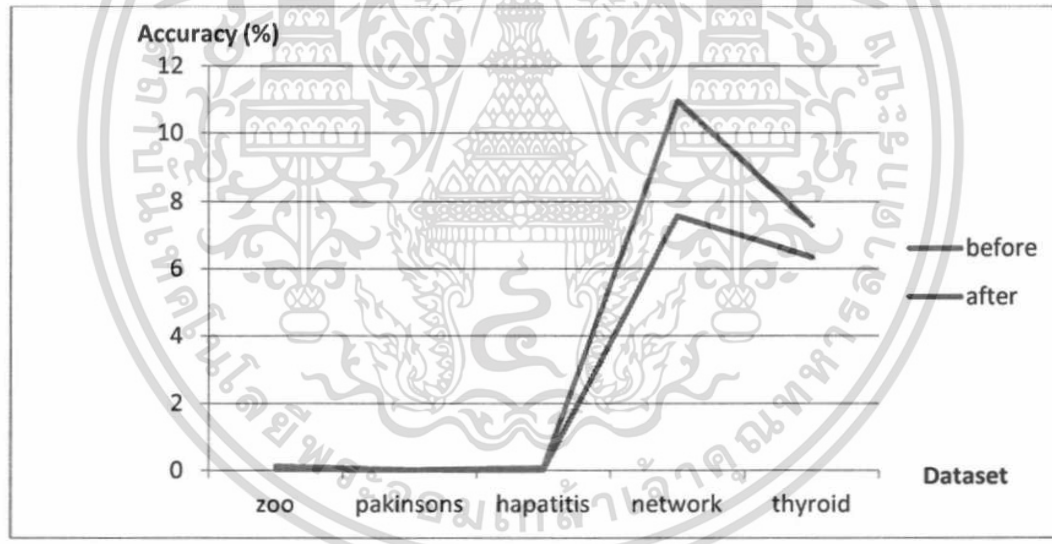
จากรูปที่ 4.9 จะเห็นว่าการคัดเลือกคุณลักษณะด้วยวิธีของกรีดดีอัลกอริธึมสามารถเพิ่มค่าความแม่นยำให้สูงขึ้น คิดเป็น 3 ใน 5 ของชุดข้อมูลที่ใช้ในการทดลอง แต่วิธีของ Information Gain และวิธีของ Principal Component Analysis สามารถเพิ่มค่าความแม่นยำให้สูงขึ้น คิดเป็น 1 ใน 5 ของชุดข้อมูลที่ใช้ในการทดลองเท่านั้น

4.4 เปรียบเทียบประสิทธิภาพและเวลาสำหรับการจำแนกข้อมูลก่อนและหลังการคัดเลือกคุณลักษณะ

ในหัวข้อนี้จะเปรียบเทียบประสิทธิภาพและเวลาที่ใช้ในการประมวลผลของการทดลองการจำแนกข้อมูลด้วย Radial basis Function โดยจะเปรียบเทียบกันสองครั้งคือ ครั้งแรกเป็นการทดลองการจำแนกข้อมูลด้วย Radial basis Function ที่มีการใช้คุณลักษณะทั้งหมดของชุดข้อมูล และครั้งที่สองเป็นการทดลองโดยการใช้กรีดดีอัลกอริธึมคัดเลือกคุณลักษณะก่อนการจำแนกข้อมูลด้วย Radial basis Function ผลการทดลอง ดังรูปที่ 4.10 และรูปที่ 4.11



รูปที่ 4.10 เปรียบเทียบประสิทธิภาพการจำแนกข้อมูลก่อนและหลังการคัดเลือกคุณลักษณะด้วย Greedy Algorithm



รูปที่ 4.11 เปรียบเทียบเวลาการประมวลผลก่อนและหลังการคัดเลือกคุณลักษณะด้วย Greedy Algorithm

จากรูปที่ 4.10 และ 4.11 แสดงให้เห็นว่าการคัดเลือกคุณลักษณะของชุดข้อมูลด้วยกรีดดีอัลกอริธึมก่อนการจำแนกข้อมูลด้วย Radial Basis Function สามารถเพิ่มค่าความแม่นยำให้กับ ชุดข้อมูลได้ อีกทั้งยังสามารถช่วยลดเวลาที่ใช้ในการประมวลผลได้ด้วย

4.5 วิเคราะห์ผลการทดลอง

ผลการทดลองคัดเลือกคุณลักษณะที่สำคัญของชุดข้อมูล โดยการใช้ของกรีดดีอัลกอริธึม ก่อนการจำแนกประเภทข้อมูลด้วย Radial Basis Function เพื่อเปรียบเทียบค่าความแม่นยำในการจำแนกประเภทข้อมูลก่อนและหลังการคัดเลือกคุณลักษณะของชุดข้อมูล ผลการเปรียบเทียบดังตารางที่ 4.4

ตารางที่ 4.1 แสดงผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลก่อนและหลังการคัดเลือกคุณลักษณะ

dataset	Before Select			After Select		
	Correctly	Incorrectly	time	Correctly	Incorrectly	time
zoo	94.06	5.94	0.12	98.02	1.98	0.03
pakinsons	84.1	15.9	0.02	87.69	12.31	0.02
hapatitis	80.65	19.35	0.08	83.87	16.13	0
network intrusion	93.5	6.5	10.95	96.6	3.4	7.58
thyroid	64.76	35.24	7.29	64.99	35.01	6.33

จากตารางที่ 4.1 ชุดข้อมูล zoo, pakinsons, hepatitis, network intrusion detection (KDD CUP99), thyroid ที่ใช้จำนวนคุณลักษณะทั้งหมดในการจำแนกข้อมูลได้ค่าความแม่นยำเท่ากับ 94.06, 84.10, 80.65, 93.50, 64.76 ตามลำดับ ค่าความแม่นยำเฉลี่ยเท่ากับ 83.41% เมื่อเปรียบเทียบกับชุดข้อมูลที่คัดเลือกคุณลักษณะแล้ว ได้ค่าความแม่นยำสูงขึ้นเท่ากับ 98.02, 87.69, 83.87, 96.6, 64.99 ตามลำดับ ค่าความแม่นยำเฉลี่ยเท่ากับ 86.23 % เวลาที่ใช้ในการประมวลผลเท่ากับ 0.12, 0.02, 0.08, 10.95, 7.29 ตามลำดับ เวลาเฉลี่ยเท่ากับ 3.69 วินาที เมื่อเปรียบเทียบกับชุดข้อมูลที่มีการคัดเลือกคุณลักษณะแล้วใช้เวลาในการประมวลผลลดลงเท่ากับ 0.03, 0.02, 0.0, 7.58, 6.33 ตามลำดับ เวลาเฉลี่ยเท่ากับ 2.79 วินาที

บทที่ 5

สรุปผลและการเสนอแนะ

5.1 สรุปผลและวิเคราะห์ผลการทดลอง

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำเสนอการประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะของชุดข้อมูลก่อนการจำแนกประเภทข้อมูล งานวิจัยนี้ได้ทำการทดลองวิธีการคัดเลือกคุณลักษณะของชุดข้อมูลด้วย 3 อัลกอริธึม คือ Greedy Algorithm, Information Gain และ Principal Component Analysis ซึ่งการคัดเลือกคุณลักษณะด้วยอัลกอริธึมที่ต่างกันจะทำให้ได้คุณลักษณะที่ต่างกันด้วย กล่าวคือ การคัดเลือกคุณลักษณะด้วย Greedy Algorithm จะคัดเลือกคุณลักษณะที่ดีที่สุด มาเป็นสถานะปัจจุบันตลอดเวลา ทำให้ได้เฉพาะคุณลักษณะที่สำคัญที่สุดและมีผลต่อค่าความแม่นยำของการจำแนกข้อมูล ส่วนการคัดเลือกคุณลักษณะของข้อมูลด้วย Information Gain และ Principal Component Analysis จะทำให้ได้คุณลักษณะที่เรียงลำดับตามค่าความสำคัญของแต่ละคุณลักษณะ เมื่อเปรียบเทียบประสิทธิภาพการคัดเลือกคุณลักษณะด้วยวิธีต่างๆ ผลการทดลองคือการคัดเลือกคุณลักษณะด้วย Greedy Algorithmสามารถเพิ่มประสิทธิภาพได้สูงที่สุด

จากนั้นนำคุณลักษณะที่ได้จากการคัดเลือกด้วยอัลกอริธึมต่างๆ ไปจำแนกประเภทข้อมูล และงานวิจัยนี้ได้ใช้อัลกอริธึมสำหรับการจำแนกประเภทข้อมูลจำนวน 3 อัลกอริธึม คือ Radial Basis Function, Multi - Layer Perceptron และ Naïve Bays โดยการจำแนกข้อมูลด้วย Multi - Layer Perceptron จะให้ค่าความแม่นยำในการจำแนกประเภทข้อมูลสูงแต่ใช้เวลาในการประมวลมาก ส่วนการจำแนกข้อมูลด้วย Naïve Bays จะให้ค่าความแม่นยำในการจำแนกข้อมูลค่อนข้างต่ำแต่ใช้เวลาในการประมวลผลน้อย ดังนั้นงานวิจัยนี้จึงพิจารณาทั้งสองด้านร่วมกันจึงเห็นว่าการจำแนกข้อมูลด้วย Radial Basis Function สามารถแก้ปัญหาทั้งสองด้านนี้ได้ กล่าวคือ วิธีนี้ให้ค่าความแม่นยำในการจำแนกข้อมูลในระดับที่ค่อนข้างสูงและใช้เวลาในการประมวลผลน้อย จากการทดลองในงานวิจัยนี้พบว่า การใช้วิธีคัดเลือกคุณลักษณะด้วย Greedy Algorithm ก่อนการจำแนกข้อมูลด้วย Radial Basis Function สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลที่สูงขึ้นและใช้เวลาประมวลผลลดลงเมื่อเปรียบเทียบกับการจำแนกข้อมูลโดยไม่มีการคัดเลือกคุณลักษณะ

5.2 ข้อเสนอแนะ

ในงานวิจัยต่อไปผู้วิจัยมีแนวคิดที่จะทดลองเปรียบเทียบค่าความแม่นยำที่ได้จากวิธีนี้กับวิธีอื่น ๆ ซึ่งให้ค่าความแม่นยำสูงแต่ใช้เวลานานเช่นวิธี Support Vector Machine เพื่อทดสอบสมมติฐานว่าวิธีที่นำเสนอจะสามารถลดเวลาในการประมวลผลแต่ยังคงให้ผลลัพธ์ที่ดีกว่าหรือดีเท่ากับวิธีที่มีอยู่ในปัจจุบัน และจะทดลองใช้วิธีการคัดเลือกคุณลักษณะด้วยวิธีอื่น ๆ เช่น Rough Set หรือ Genetic Algorithm เพื่อทดสอบว่าวิธีการดังกล่าวจะสามารถเพิ่มความแม่นยำให้กับวิธีที่นำเสนอได้หรือไม่

เอกสารอ้างอิง

[1] Jing Bi, Kun Zhang, Xiaojing Cheng, 2009, "Intrusion Detection Base on RBF Neural Network", International Symposium on Information Engineering and Electronic Commerce, 362-365.

[2] Lin Li-zhong, Liu Zhi-guo, Duan Xian-hui, 2010, "Network Intrusion Detection by a Hybrid Method of Rough Set and RBF Neural Network", 2nd International Conference on Education Technology and Computer (ICETC), v3-317 - v3-320.

[3] Thammarath Pratchayawasin, Veera Boonjing, 2011, "A Gain Positive Region Reduct Selection for Back Propagation Neural Network Classification", 3rd Conference on Knowledge and Smart Technologies, 51-57.

[4] พลอยพรรณ สอนสุวิทย์, ตรัสพงศ์ ไทยอุปถัมภ์, 2009, "การเปรียบเทียบประสิทธิภาพการตรวจจับสิ่งผิดปกติทางเครือข่ายชนิด Probing", 5th National Conference on Computing and Information Technology, 425-430.

[5] วงกต ศรีอุไร, พยุง มีสัจ, ชูชาติ หฤไชยะศักดิ์, 2009, "การเตรียมฟีเจอร์บนพื้นฐานแบบจำลองหัวข้อสำหรับการจำแนกหมวดหมู่ของเอกสาร", 5th National Conference on Computing and Information Technology, 146-151.

[6] ภัทราวุฒิ แสงศิริ, ศจีมาจ ณ วิเชียร, 2010, "การคัดแยกประเภทของมะเร็งเม็ดเลือดขาว โดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน", 11th Graduate Research Conference, Khon Kaen University, SDO1-1 – SDO1-9.

[7] กิตติพล วิแสง, สิรภัทร เชี่ยวชาญวัฒนา, คำรณ สุนัติ,2009 ,"การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน", 5th National Conference on Computing and Information Technology.

เอกสารอ้างอิง (ต่อ)

[8] ภรัณยา อำมฤครัตน์, พยุง มีสัจ, 2010, "การเปรียบเทียบประสิทธิภาพการลดมิติข้อมูลและจำแนกข้อมูลโดยวิธีการทางเครือข่ายประสาท", 11th Graduate Research Conference, Khon Kaen University, 58-65.

[9] ธรรมศักดิ์ เธียรนิเวศ, 2008, "การลดขนาดข้อมูลด้วยน้ำหนักความหนาแน่นเพื่อการจัดกลุ่มข้อมูลขนาดใหญ่", วิทยานิพนธ์ สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

[10] http://archive.ics.uci.edu/ml/datasets.html

[11] http://www.cs.waikato.ac.nz/ml/weka/

[12] สิริภัทร เชี่ยงชาญวัฒนา, 2009, "ข่ายงานเรเดียลเบสิสฟังก์ชั่น (Radial Basis Function)", เอกสารประกอบการสอนวิชาข่ายงานประสาทเทียม (Artificial Neural Network) ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น

ภาคผนวก ก

งานวิจัยที่ตีพิมพ์

การจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่สำคัญ

Data Classification by Selecting Important Attributes

จิราภรณ์ ถมแก้ว* ศรัณย์ อินทโกสุม

สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

กรุงเทพมหานคร 10520

**E-mail: jira_1629_@hotmail.com*

บทคัดย่อ

การจำแนกข้อมูลจากงานวิจัยที่ผ่านมามักจะพิจารณาคุณลักษณะทั้งหมดของข้อมูล อย่างไรก็ตามคุณลักษณะบางประการมีความสำคัญน้อยซึ่งเมื่อนำมาร่วมคำนวณด้วยแล้วอาจเป็นสาเหตุทำให้ความแม่นยำในการจำแนกข้อมูลลดลง งานวิจัยนี้ทดสอบสมมติฐานดังกล่าวโดยการประยุกต์ใช้กรีดดีอัลกอริธึมเพื่อคัดเลือกคุณลักษณะที่สำคัญของข้อมูล ร่วมกับการจำแนกข้อมูลซึ่งผู้วิจัยเลือกใช้เรเดียลเบสิสฟังก์ชัน เหตุผลหลักในการเลือกก็คือวิธีดังกล่าวให้ค่าความแม่นยำในระดับที่ยอมรับได้ และใช้เวลาในการทำงานที่เหมาะสม ผลการทดลองพบว่าวิธีการที่นำเสนอเพิ่มความแม่นยำของการจำแนกข้อมูลและลดเวลาที่ใช้ในการประมวลผล

คำสำคัญ: การคัดเลือกคุณลักษณะ, กรีดดีอัลกอริธึม, เรเดียลเบสิสฟังก์ชัน

บทนำ

การจำแนกข้อมูล [8] เป็นเทคนิคหนึ่งในการทำเหมืองข้อมูล โดยการจำแนกข้อมูลให้อยู่ในกลุ่มเดียวกันตามที่กำหนดโดยการสร้างกฎเพื่อช่วยการตัดสินใจจากข้อมูลที่มีอยู่ และมีการนำมาประยุกต์ใช้กับงานด้านต่างๆ เช่น การจำแนกผู้ป่วยโรคต่างๆ การจำแนกกลุ่มลูกค้า การจำแนกเอกสารสารสนเทศ เป็นต้น ซึ่งการจำแนกข้อมูลดังกล่าวต้องใช้คุณลักษณะของข้อมูลเป็นสำคัญ

คุณลักษณะของข้อมูล คือ ลักษณะหรือคุณสมบัติที่ใช้ระบุองค์ประกอบหรือรายละเอียดของชุดข้อมูล คุณลักษณะของข้อมูลที่ดีต้องมีความถูกต้องและเชื่อถือได้ เก็บเฉพาะข้อมูลที่จำเป็นต้องใช้และครบถ้วนสมบูรณ์ อย่างไรก็ตามหากชุดข้อมูลมีการเก็บคุณลักษณะที่มากเกินความจำเป็นจะทำให้สิ้นเปลืองทรัพยากรและเวลา การนำข้อมูลไปใช้งานอาจทำให้ประสิทธิภาพลดลงได้ การคัดเลือกคุณลักษณะของชุดข้อมูล [1] ก่อนจำแนกข้อมูลเป็นกระบวนการเพื่อลดขนาดมิติของข้อมูลเดิม แต่ทำให้สูญเสียความสำคัญของข้อมูลน้อยที่สุด [9]

งานวิจัยนี้เสนอวิธีการคัดเลือกเฉพาะคุณลักษณะที่สำคัญที่สุดและมีผลต่อค่าความถูกต้องข้อมูล เพื่อเพิ่มค่าความแม่นยำ (Accuracy) และลดเวลาในการประมวลผล โดยโครงสร้างส่วนที่เหลือของบทความเป็นดังนี้ ส่วนที่สองกล่าวถึงงานวิจัยและทฤษฎีที่เกี่ยวข้อง ส่วนที่สามวิธีการทดลอง ผลการทดลองอยู่ในส่วนที่สี่ และส่วนที่ห้าคือส่วนสรุปและข้อเสนอแนะ

งานวิจัยและทฤษฎีที่เกี่ยวข้อง

1 การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะ คือการลดขนาดของข้อมูลโดยการทำให้ข้อมูลเดิมมีขนาดลดลงและสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด เทคนิคการคัดเลือกที่ต่างกันทำให้ได้คุณลักษณะที่สำคัญต่างกันด้วย การคัดเลือกคุณลักษณะของชุดข้อมูล มีงานวิจัยที่เกี่ยวกับการคัดเลือกคุณลักษณะดังนี้

พลอยพรรณ สอนสุวิทย์ [4] นำเสนอวิธีตรวจจับการบุกรุกเครือข่ายชนิด Probing ที่มีคุณลักษณะจำนวนมาก จึงใช้ Genetic Algorithm ในการคัดเลือกคุณลักษณะจากการ

ทดลองพบว่าการคัดเลือกคุณลักษณะที่ไม่สำคัญออก ไม่มีผลต่อค่าความถูกต้องของข้อมูลและยังสามารถลดเวลาการทำงานได้ด้วย

วงกต ศรีอุไร [5] ปรับปรุงการจำแนกหมวดหมู่ของเอกสารโดยวิธีสร้างแบบจำลองหัวข้อให้กับเอกสารและใช้วิธีการคัดเลือกคุณลักษณะสองวิธี คือ Information Gain และ Chi Squared จากการทดลองพบว่าการคัดเลือกคุณลักษณะด้วยวิธี Information Gain ให้ประสิทธิภาพดีกว่าความแม่นยำเพิ่มขึ้น 10.2%

จิราพร สุดใหญ่ [10] นำเสนอการคัดเลือกคุณลักษณะข้อมูลเพื่อลดมิติของตัวแปรที่ไม่มีความสัมพันธ์กัน ขั้นตอนวิธีที่นำมาใช้เพื่อลดมิติข้อมูลคือ Principal Component Analysis, Linear Discriminate Analysis, Maximum margin criterion โดยเลือกตัดมิติค่าความสมนัยที่มีค่าน้อย ผลการทดลองพบว่าการคัดเลือกด้วย Linear Discriminate Analysis ให้ค่าความแม่นยำที่สูงกว่า

นอกจากงานวิจัยที่กล่าวมาแล้วยังมีเทคนิคการคัดเลือกคุณลักษณะอื่นๆ เช่น

1.1 Rough Set

รัฟเซ็ต เป็นทฤษฎีที่ใช้จัดการเกี่ยวกับความคลุมเครือและความไม่แน่นอนของข้อมูล [2, 3] จากทฤษฎีนี้จะได้ผลลัพธ์เป็นข้อมูล 2 อย่างคือ การประมาณขอบเขตล่าง คือชุดข้อมูลที่ไม่มีความคลุมเครือ และการประมาณขอบเขตบน คือชุดข้อมูลที่มีความน่าจะเป็นที่จะเป็นข้อมูลในชุดแรก

1.2 Information Gain

เป็นเทคนิคการคัดเลือกมิติของข้อมูล [5][6] โดยการวัดค่า Gain ของแต่ละโหนด ถ้าโหนดใดมีค่า Gain สูงสุดก็จะถูกเลือกเป็นโหนดราก และนำข้อมูลที่เหลือมาหาค่า Gain อีกครั้งเพื่อให้ได้โหนดต่อไป คำนวณได้จาก

$$Gain(S,A) = E(S) - \sum_{v \in Value(A)} \frac{|S_v|}{S} E(S_v) \quad (1)$$

เมื่อ Gain(S,A) = ค่า Gain ของเหตุการณ์ที่สนใจ E(S) = ค่า Entrophy ของเซ็ตข้อมูลที่สนใจ S_v = เซ็ตข้อมูลที่สนใจ

1.3 Greedy Algorithm

กรีดดีอัลกอริธึม เป็นการค้นหาแบบไปข้างหน้า และค้นหาแบบดีที่สุดก่อน (Best First Search) โดยการสร้างกราฟต้นไม้ จากนั้นจะค้นหาแบบชั้นต่อชั้น ระหว่างโหนดจะเชื่อมต่อกันด้วยกิ่งที่มีค่าน้ำหนักของแต่ละกิ่ง คำนวณได้จาก

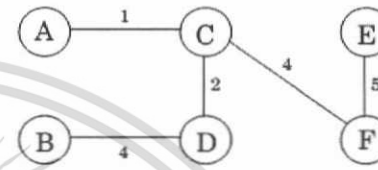
Input: G = (V;E)

เมื่อ V = จำนวนโหนด, E = จำนวนกิ่ง, Output: T = (V; E'), เมื่อ $E' \subseteq E$

Weight (T) = $\sum w_e$

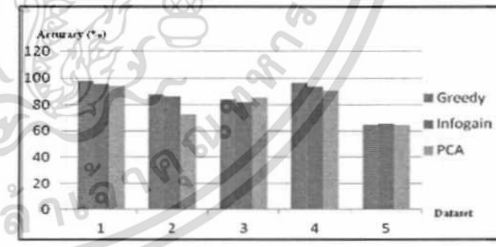
เมื่อ w_e = น้ำหนักแต่ละกิ่ง

เทคนิคของกรีดดีอัลกอริธึม จะพิจารณาเลือกทางเลือกที่สามารถเชื่อมต่อกันได้ทุกโหนด แต่ไม่ก่อให้เกิดเป็นกราฟวงกลม และมีค่าน้ำหนักรวมของทุกโหนดน้อยที่สุด



รูปที่ 1 การทำงานของกรีดดีอัลกอริธึม

การคัดเลือกคุณลักษณะด้วยทฤษฎีที่กล่าวมาส่วนใหญ่จะให้ค่า Rank ของคุณลักษณะแต่ละตัว การนำไปใช้ต้องกำหนดเกณฑ์การเลือกคุณลักษณะเอง ซึ่งอาจทำให้ค่าผิดพลาดได้ แต่เทคนิคกรีดดีอัลกอริธึม จะได้เฉพาะคุณลักษณะที่สำคัญเท่านั้นและสามารถนำไปใช้ได้ทันที และเมื่อนำคุณลักษณะที่ได้ไปจำแนกประเภทข้อมูล มีค่าความแม่นยำที่สูงกว่า ดังรูปที่ 2 ผู้วิจัยจึงเลือกใช้เทคนิคของกรีดดีอัลกอริธึม



รูปที่ 2 เปรียบเทียบผลการคัดเลือกคุณลักษณะ

2 การจำแนกข้อมูล (Data Classification)

การจำแนกข้อมูล คือการจำแนกข้อมูลแบบรู้ล่วงหน้าว่าข้อมูลมีกี่ประเภทและประเภทอะไรบ้าง หรือเรียกว่าการเรียนรู้แบบมีผู้สอน(Supervised learning) ซึ่งต้องการผู้สอนเพื่อให้ความรู้กับข้อมูลก่อนว่าประเภทข้อมูลที่ต้องการมีกี่ประเภท ข้อมูลมีลักษณะอย่างไร เพื่อนำการเรียนรู้นั้นไปใช้สำหรับการตัดสินใจ เทคนิคที่ใช้สำหรับการจำแนกข้อมูลมีให้เลือกใช้หลายเทคนิค เช่น

2.1 Naïve Bays

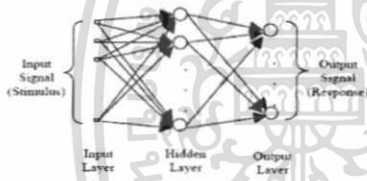
Naïve Bayes [7] เป็นเทคนิคที่ใช้ทฤษฎีของ Bayes Theorem โดยมีความน่าจะเป็นที่จะเกิดเหตุการณ์ต่างๆ ที่ใช้ในการจัดกลุ่มนั้นเป็นอิสระต่อกัน ซึ่งจะวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตาม เพื่อใช้สร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ โดยคำนวณได้จาก

$$P(H|E) = [P(E|H) \times P(H)] / P(E)$$

เมื่อ P(H) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ H และ P(H|E) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ H เมื่อเกิดเหตุการณ์ E

2.2 Multilayer Perceptron (MLP)

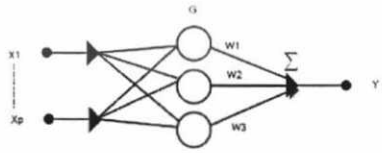
Multilayer Perceptron [4] เป็นโครงข่ายประสาทเทียมแบบหลายชั้น โหนดในแต่ละชั้นจะมีค่าน้ำหนัก (weight) ที่แตกต่างกัน เทคนิคนี้ให้ค่าความแม่นยำจากการคำนวณสูง แต่ใช้เวลาในการประมวลผลนาน เหมาะกับงานที่มีความซับซ้อนสูง แสดงโครงสร้างดังรูปที่ 3



รูปที่ 3 แสดงโครงข่าย MLP

2.3 Radial Basis Function (RBF)

Radial Basis Function (RBF) [1][2] เป็นโครงข่ายสำหรับการจำแนกข้อมูลแบบป้อนไปข้างหน้าโดยการคำนวณหาจุดศูนย์กลางของข้อมูลและแบ่งข้อมูลออกเป็นกลุ่มเพื่อจัดกลุ่มของข้อมูล โครงข่าย RBFประกอบด้วย 3 ชั้น คือ ชั้นข้อมูลเข้า (Input layer) ชั้นซ่อน (Hidden layer) และชั้นข้อมูลออก (Output Layer) ดังรูปที่ 4



รูปที่ 4 แสดงโครงข่าย RBF

จากรูปที่ 3 เมื่อ Xi คือ ข้อมูลเข้า (Input) ที่ถูกส่งค่าไปยังชั้นซ่อน (Hidden Layer) ซึ่งคำนวณได้จาก

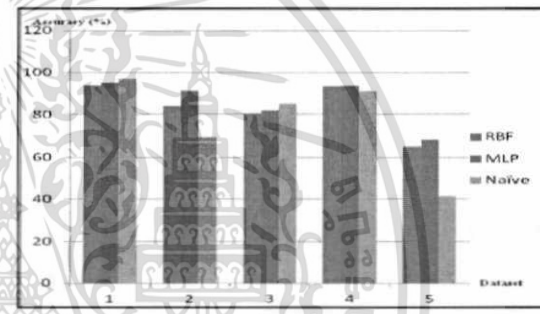
$$\theta_j(x) = \exp = \frac{x - c_j e}{2\sigma^2 j} \quad \text{เมื่อ } j = 1,2,\ldots,n \qquad (2)$$

เมื่อ θ คือข้อมูลออกที่ j ในชั้นซ่อน, x คือข้อมูลเข้า, C_j คือ ศูนย์กลางของโหนดที่ j จากนั้นคำนวณค่าข้อมูลออก (Output) ซึ่งคำนวณได้จาก

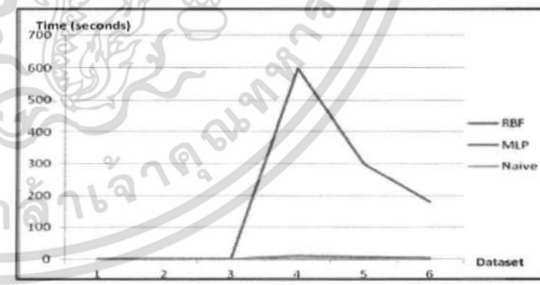
$$y = i_c(k+1) = \sum_{i=1}^{n} w_i \theta_i(x) \qquad (3)$$

เมื่อ W_j คือค่าน้ำหนักระหว่างชั้นซ่อนและชั้นข้อมูลออก และ y คือผลลัพธ์

งานวิจัยที่ผ่านมาพบว่าเทคนิค Naïve Bayes ใช้เวลาการทำงานน้อย แต่ให้ค่าความแม่นยำต่ำ เทคนิค MLP ใช้เวลาการทำงานมาก แต่ให้ค่าความแม่นยำสูง ส่วนเทคนิค RBF ให้ค่าความแม่นยำสูง และใช้เวลาการทำงานน้อย ดังรูปที่ 5 และรูปที่ 6



รูปที่ 5 เปรียบเทียบประสิทธิภาพการทำงานของ RBF, MLP, Naïve bayes



รูปที่ 6 เปรียบเทียบเวลาการทำงานของ RBF, MLP, Naïve bayes

งานวิจัยนี้จึงพิจารณาทั้งสองด้านร่วมกันและพบว่าเทคนิค RBF จะสามารถแก้ปัญหาดังกล่าวได้

วิธีการทดลอง

การทดลองนี้ใช้โปรแกรม weka 3.6 ในการทดลอง [11] ซึ่งเป็นโปรแกรมที่ใช้ในการทำเหมืองข้อมูล (Data Mining) ที่ได้รับความนิยมและใช้งานกันอย่างแพร่หลาย และได้นำชุดข้อมูลจาก UCI Machine Learning Repository [10]

จำนวน 5 ชุด ได้แก่ zoo, pakinsons, hepatitis, network intrusion detection (KDD CUP99), thyroid ข้อมูลแต่ละชุดจะมีจำนวนคุณลักษณะ ชนิดของคุณลักษณะ และจำนวนแถวข้อมูลที่แตกต่างกัน ดังตารางที่ 1

ตารางที่ 1 ตัวอย่างชุดข้อมูลที่นำมาทดสอบ

dataset	attributes	instances	attribute type
zoo	17	101	Integer, Nominal
pakinsons	23	197	Real
hapatitis	20	155	Integer, Real, Nominal
network intrusion	42	1000	Integer,Nominal
thyroid	29	9172	Integer, Real, Nominal

จากตารางที่ 1 ชุดข้อมูล zoo, pakinsons, hepatitis, network intrusion detection (KDD CUP99), thyroid มีจำนวนคุณลักษณะเท่ากับ 17, 23, 20, 42, 29 ตามลำดับ จากนั้นนำชุดข้อมูลแต่ละชุดไปคัดเลือกคุณลักษณะของชุดข้อมูลด้วย กรีดดีอัลกอริธึมของโปรแกรม weka โดยเรียกใช้ได้จาก weka/attributeSelection/GreedyStepwise ผลจากการคัดเลือกคุณลักษณะของข้อมูลแต่ละชุด ดังตารางที่ 2

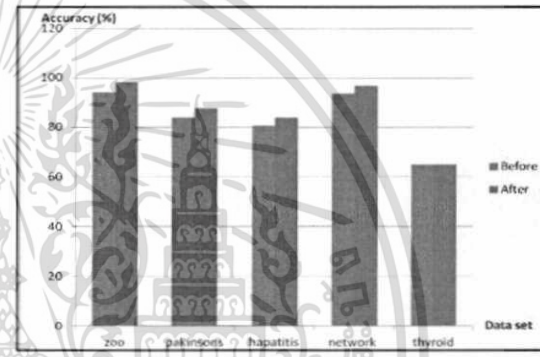
ตารางที่ 2 จำนวนคุณลักษณะจากการคัดเลือก ด้วย กรีดดีอัลกอริธึม

dataset	attributes
zoo	9
pakinsons	10
hapatitis	10
network intrusion	6
thyroid	11

จากตารางที่ 2 เมื่อคัดเลือกคุณลักษณะด้วยกรีดดีอัลกอริธึมแล้วเหลือจำนวนคุณลักษณะลดลงเท่ากับ 9, 10, 10, 6, 11 ตามลำดับ จากนั้นนำชุดข้อมูลที่ได้จากการคัดเลือกคุณลักษณะดังกล่าวไปจำแนกข้อมูลด้วย Radial Basis Function (RBF) งานวิจัยนี้ใช้การทดลองแบบ 10-fold cross-validation เพื่อแบ่งชุดข้อมูลสำหรับการสอนและการทดสอบ

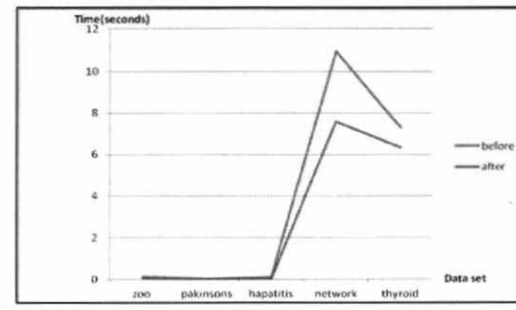
ผลการศึกษาวิจัยและอภิปรายผล

การคัดเลือกคุณลักษณะของชุดข้อมูลเพื่อลดมิติของข้อมูลก่อนการจำแนกข้อมูล โดยการใช้กรีดดีอัลกอริธึม ทำให้ได้เฉพาะคุณลักษณะที่มีความสำคัญต่อค่าความแม่นยำของชุดข้อมูลเท่านั้นและสามารถนำไปใช้ได้ทันที เมื่อนำคุณลักษณะที่ได้มาจำแนกข้อมูลด้วย RBF ทำให้ได้ค่าความแม่นยำที่สูงขึ้น ผลการเปรียบเทียบดังรูปที่ 7



รูปที่ 7 เปรียบเทียบผลการทดลอง

จากรูปที่ 7 ชุดข้อมูล zoo, pakinsons, hepatitis, network intrusion detection (KDD CUP99), thyroid ที่ใช้จำนวนคุณลักษณะทั้งหมดในการจำแนกข้อมูลได้ค่าความแม่นยำเท่ากับ 94.06, 84.10, 80.65, 93.50, 64.76 ตามลำดับ ค่าความแม่นยำเฉลี่ยเท่ากับ 83.41% เมื่อเปรียบเทียบกับชุดข้อมูลที่คัดเลือกคุณลักษณะแล้ว ได้ค่าความแม่นยำสูงขึ้นเท่ากับ 98.02, 87.69, 83.87, 96.6, 64.99 ตามลำดับ ค่าความแม่นยำเฉลี่ยเท่ากับ 86.23 % เวลาที่ใช้ในการประมวล ดังรูปที่ 8



รูปที่ 8 เปรียบเทียบเวลาที่ใช้ในการทดลอง

จากรูปที่ 8 ชุดข้อมูลที่ใช้คุณลักษณะทั้งหมดใช้เวลาในการประมวลผลเท่ากับ 0.12, 0.02, 0.08, 10.95, 7.29 ตามลำดับ เวลาเฉลี่ยเท่ากับ 3.69 วินาที เมื่อเปรียบเทียบกับชุดข้อมูลที่มีการคัดเลือกคุณลักษณะแล้วใช้เวลาในการประมวลผลลดลงเท่ากับ 0.03, 0.02, 0.0, 7.58, 6.33 ตามลำดับ เวลาเฉลี่ยเท่ากับ 2.79 วินาที

สรุปและข้อเสนอแนะ

งานวิจัยนี้เสนอการประยุกต์ใช้ วิธีการคัดเลือกคุณลักษณะของชุดข้อมูลด้วยกรีดดีอัลกอริธึมก่อนการจำแนกข้อมูล ซึ่งวิธีนี้จะคัดเลือกคุณลักษณะที่สำคัญที่สุดและมีผลต่อค่าความแม่นยำของการจำแนกข้อมูล ผลการทดลองพบว่าการใช้วิธีคัดเลือกคุณลักษณะด้วยกรีดดีอัลกอริธึมร่วมกับการจำแนกข้อมูลด้วย RBF สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลให้สูงขึ้นและใช้เวลาประมวลผลลดลงเมื่อเปรียบเทียบกับการจำแนกข้อมูลโดยไม่มีการคัดเลือกคุณลักษณะ จากผลการทดลองงานวิจัยครั้งนี้ สามารถนำมาเป็นเครื่องมือช่วยวิเคราะห์และตัดสินใจเกี่ยวการจำแนกข้อมูลต่างๆ ได้ เช่น ข้อมูลลูกค้า ข้อมูลผู้ป่วย ข้อมูลสารสนเทศ เป็นต้น ให้มีความแม่นยำมากขึ้น

ในงานวิจัยต่อไปผู้วิจัยมีแนวคิดที่จะทดลองใช้วิธีการคัดเลือกคุณลักษณะด้วยวิธีอื่น ๆ เช่น รัฟเซ็ต เพื่อทดสอบว่าวิธีการดังกล่าวจะสามารถเพิ่มความแม่นยำให้กับวิธีที่นำเสนอได้หรือไม่

เอกสารอ้างอิง

[1] Jing Bi, Kun Zhang, Xiaojing Cheng, 2009, "Intrusion Detection Base on RBF Neural Network", International Symposium on Information Engineering and Electronic Commerce, 362-365.

[2] Lin Li-zhong, Liu Zhi-guo, Duan Xian-hui, 2010, "Network Intrusion Detection by a Hybrid Method of Rough Set and RBF Neural Network", 2nd International Conference on Education Technology and Computer (ICETC), v3-317 - v3-320.

[3] Thammarath Pratchayawasin, Veera Boonjing, 2011, "A Gain Positive Region Reduct Selection for Back Propagation Neural Network Classification", 3rd Conference on Knowledge and Smart Technologies, 51-57.

[4] พลอยพรรณ สอนสุวิทย์, ตรัสพงศ์ ไทยอุปถัมภ์, 2009, "การเปรียบเทียบประสิทธิภาพการตรวจจับสิ่งผิดปกติทางเครือข่ายชนิด Probing", 5th National Conference on Computing and Information Technology, 425-430.

[5] วงกต ศรีอุไร, พยุง มีสัจ, ชูชาติ หฤไชยะศักดิ์, 2009, "การเตรียมฟีเจอร์บนพื้นฐานแบบจำลองหัวข้อสำหรับการจำแนกหมวดหมู่ของเอกสาร", 5th National Conference on Computing and Information Technology, 146-151.

[6] ภัทราวุฒิ แสงศิริ, ศจีมาจ ณ วิเชียร, 2010, "การคัดแยกประเภทของมะเร็งเม็ดเลือดขาว โดยใช้วิธีการจัดอันดับร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน", 11th Graduate Research Conference, Khon Kaen University, SDO1-1 – SDO1-9.

[7] กิตติพล วิแสง, สิรภัทร เชี่ยวชาญวัฒนา, คำรณ สุนัติ ,2009 ,"การวิเคราะห์ปัจจัยเสี่ยงของโรคเบาหวาน", 5th National Conference on Computing and Information Technology.

[8] ภรัณยา อำมฤครัตน์, พยุง มีสัจ, 2010, "การเปรียบเทียบประสิทธิภาพการลดมิติข้อมูลและจำแนกข้อมูลโดยวิธีการทางเครือข่ายประสาท", 11th Graduate Research Conference, Khon Kaen University, 58-65.

[9] ธรรมศักดิ์ เธียรนิเวศ, 2008, "การลดขนาดข้อมูลด้วยน้ำหนักความหนาแน่นเพื่อการจัดกลุ่มข้อมูลขนาดใหญ่", วิทยานิพนธ์ สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

[10] จิราพร สุดใหญ่, สิริภัทร เชี่ยวชาญวัฒนา, คำรณ สุนัติ, 2010,"การคัดเลือกคุณลักษณะข้อมูลสำหรับปัญหาการจำแนกประเภทข้อมูล ด้วยเทคนิคโครงข่ายประสาทเทียมแบบเอ็กทรีม", The 14th National Computer Science and Engineering Conference, 89-94.

[11] http://archive.ics.uci.edu/ml/datasets.html

[12] http://www.cs.waikato.ac.nz/ml/weka/

ประวัติผู้เขียน

ชื่อ – สกุล	นางสาวจิราภรณ์ ถมแก้ว
วัน เดือน ปีเกิด	1 มิถุนายน 2529
ที่อยู่	41/2 หมู่ที่ 3 ต.ถ้ำใหญ่ อ.ทุ่งสง จ.นครศรีธรรมราช 80110
E-mail	jira_1629_@hotmail.com
ประวัติการศึกษา	2551 จบการศึกษาปริญญาบริหารธุรกิจบัณฑิต สาขาระบบสารสนเทศทางคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย วิทยาเขตนครศรีธรรมราช