การเลือกคุณลักษณะโดยใช้อัตราส่วนเกนสำหรับการจำแนกประเภทข้อมูลด้วยขั้นตอนวิธีเชิงพันธุกรรม

โดย นายบุญญาพล เข็มปัญญา

A GAIN RATIO BASED FEATURE SELECTION FOR CLASSIFICATION USING GENETIC ALGORITHM

บุญญาพล เข็มปัญญา
BOONYAPAWN KEMPANYA

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาการจัดการสารสนเทศ

คณะวิทยาการสารสนเทศ

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

พ.ศ. 2555

ลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

การคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภท โดยใช้ขั้นตอนวิธีเจเนติก

A GAIN RATIO BASED FEATURE SELECTION FOR CLASSIFICATION USING GENETIC ALGORITHM

บุญญาพร เข็มปัญญา

BOONYAPAWN KHEMPANYA

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2555

KMITL-2012-SC-M-002-010

A GAIN RATIO BASED FEATURE SELECTION FOR CLASSIFICATION USING GENETIC ALGORITHM

BOONYAPAWN KHEMPANYA

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-002-010

คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก
A Gain Ratio Based Feature Selection for Classification using Genetic Algorithm

นักศึกษา นางสาวบุญญาพร เข็มปัญญา
รหัสประจำตัว 51609252
ปริญญา วิทยาศาสตรมหาบัณฑิต
สาขาวิชา วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ดร.วีระ บุญจริง

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ดร.อนันตพร	ศรีสวัสดิ์	อนันตพร ศรีสวัสดิ์
ผศ.ดร.ศรัณย์	อินทโกสุม	
ดร.เฉลิมศักดิ์	เลิศวงศ์เสถียร	
รศ.ดร.วีระ	บุญจริง	

วัน / เดือน/ ปี ที่สอบ 19 เมษายน พ.ศ. 2555 เวลา 9.00 – 12.00 น.
สถานที่สอบ ณ ห้อง 216 ชั้น 2 อาคารจุฬาภรณวลัยลักษณ์ 1

คณะวิทยาศาสตร์รับรองแล้ว

(รองศาสตราจารย์ ดร.ดุษฎี ธนะบริพัฒน์)
คณบดีคณะวิทยาศาสตร์
วันที่ 26 **เดือน** มิถุนายน **พ.ศ.** 55

หัวข้อวิทยานิพนธ์	การคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก
นักศึกษา	นางสาวบุญญาพร เข็มปัญญา
รหัสประจำตัว	51609252
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2555
อาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง

บทคัดย่อ

งานวิจัยนี้เสนอวิธีการคัดเลือกกลุ่มลักษณะโดยพิจารณาค่าอัตราส่วนเกนของแต่ละลักษณะ กลุ่มลักษณะที่ถูกคัดเลือกจะเป็นกลุ่มที่มีค่าอัตราส่วนเกนเฉลี่ยสูงสุด เนื่องจาก การค้นหากลุ่มดังกล่าวต้องทำการประเมินกลุ่มย่อยที่เป็นไปได้ทั้งหมดซึ่งในปัญหาที่มีลักษณะจำนวนมากทำได้ยาก งานวิจัยนี้จึงใช้ขั้นตอนวิธีเจเนติกในการค้นหากลุ่มลักษณะดังกล่าว เมื่อนำกลุ่มลักษณะที่ได้มาทดสอบกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับเพื่อจำแนกประเภทข้อมูล พบว่าโครงข่ายนี้ให้ความแม่นยำในการจำแนกประเภทสูงกว่าโครงข่ายที่สร้างจากลักษณะทั้งหมดอย่างมีนัยสำคัญ

คำสำคัญ: การคัดเลือกลักษณะ, เจเนติกอัลกอริทึม, ค่าอัตราส่วนเกน

Thesis Title	A Gain Ratio Based Feature Selection for Classification using Genetic Algorithm
Student	Boonyapawn Khempanya
Student ID	51609252
Degree	Master of Science
Program	Computer Science
Year	2012
Thesis Advisor	Assoc.Prof.Dr.Veera Boonjing

ABSTRACT

This research proposes a method for selecting a set of attributes by using a gain ratio of each attribute. The selected set of attributes will be the set the highest average rate of the gain ratio. The method of selection requires an evaluation of all possible sets which is difficult for the problem with a large number of attributes. This research therefore uses a genetic algorithm to select the set of attributes. When set of attributes selected by this method, was tested with backpropagation neural network for classification, it is found that this network sives accuracy of classification much higher than those networks learned with all attributes.

Keywords: Feature Selection, Genetic Algorithm, Gain Ratio

กิตติกรรมประกาศ

วิทยานิพนธ์นี้มิอาจสำเร็จลุล่วงไปได้ด้วยดี หากมิได้รับคำแนะนำ คำชี้แจง ความรู้และความเอาใจใส่จาก รศ.ดร.วีระ บุญจริง ผู้เป็นอาจารย์ที่ปรึกษา ซึ่งท่านได้สละเวลาให้กับข้าพเจ้าอย่างเต็มที่จึงใคร่ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ผศ.ดร.ศรัณย์ อินทโกสุม ดร.อนันตพร ศรีสวัสดิ์ และดร.เฉลิมศักดิ์ เลิศวงศ์เสถียร คณะกรรมการสอบหัวข้อ และโครงร่างวิทยานิพนธ์ที่กรุณาให้คำแนะนำตลอดจนข้อชี้แนะจนในที่สุดทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลงได้

ขอขอบพระคุณบิดามารดา และพี่ๆที่ให้การสนับสนุนให้ได้เรียนในระดับที่ได้ตั้งใจ อีกทั้งยังได้ดูแลเรื่องค่าใช้จ่ายต่างๆระหว่างการศึกษาเป็นอย่างดี

ขอขอบคุณพี่ๆและเพื่อนๆทุกคนโดยเฉพาะ นายมานิตย์ ขวัญยืนที่ให้คำปรึกษาและอำนวยความสะดวกในด้านต่างๆ

สำหรับคุณงามความดีและประโยชน์อันใดที่เกิดขึ้นจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบให้กับบิดา มารดา อาจารย์ทุกท่านซึ่งเป็นที่เคารพรักยิ่ง ตลอดจนญาติพี่น้อง และเพื่อนๆทุกคน

บุญญาพร เข็มปัญญา

สารบัญ

หน้า

สารบัญ(ต่อ)

สารบัญตาราง

สารบัญรูป

รูปที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การคัดเลือกลักษณะ เป็นขั้นตอนหนึ่งในการเตรียมข้อมูลสำหรับการทำเหมืองข้อมูล เพื่อที่จะหากลุ่มลักษณะที่ดีที่สุดและลดข้อมูลไม่ตรงประเด็น หรือข้อมูลซ้ำซ้อน ที่ส่งผลต่อประสิทธิภาพและความถูกต้องในการวิเคราะห์ข้อมูล ดังนั้นการคัดเลือกลักษณะที่ดีจะส่งผลให้ประสิทธิภาพของอัลกอริทึมในการวิเคราะห์ข้อมูลหรือจำแนกประเภทข้อมูลมีความแม่นยำสูงขึ้น

ปัญหาการคัดเลือกลักษณะสำหรับการจำแนกประเภทข้อมูลเป็นปัญหาการค้นหากลุ่มลักษณะย่อยที่ดีที่สุดตามเกณฑ์ที่กำหนดโดยเกณฑ์ที่ใช้ต้องสามารถวัดความสำคัญของแต่ละลักษณะได้ อัตราส่วนเกน (Gain Ratio) เป็นเกณฑ์ที่มีการขจัดความเอนเอียงของลักษณะทำให้การประเมินมีความถูกต้องสูงขึ้น งานวิจัยนี้จึงเลือกใช้เกณฑ์ดังกล่าวในการวัดความสำคัญของแต่ละลักษณะต่องานจำแนกประเภท สำหรับการค้นหากลุ่มลักษณะที่ดีที่สุดตามเกณฑ์นั้นต้องมีการประเมินแต่ละกลุ่มจากกลุ่มย่อยที่เป็นไปได้ทั้งหมด ซึ่งในปัญหาที่มีลักษณะเป็นจำนวนมากจะทำได้ยาก เนื่องจากมีกลุ่มย่อยที่ต้องประเมินจำนวนมหาศาลเมื่อเทียบกับจำนวนลักษณะ ดังนั้น การค้นหาคำตอบจึงต้องทำโดยใช้ขั้นตอนวิธีการประมาณ งานวิจัยนี้จึงเสนอใช้ขั้นตอนวิธีการเจเนติกในการค้นหาคำตอบดังกล่าว โดยใช้ค่าอัตราส่วนเกนเฉลี่ยเป็นฟังก์ชันประเมินความเหมาะสมเพื่อให้ได้กลุ่มลักษณะที่ดีที่สุดสำหรับการจำแนกประเภท

1.2 วัตถุประสงค์

งานวิจัยนี้มีวัตถุประสงค์เพื่อประยุกต์วิธีการค้นหากลุ่มลักษณะที่ดีที่สุดโดยลดการประเมินกลุ่มลักษณะย่อยที่เป็นไปได้ทั้งหมด แล้วได้ความแม่นยำในการจำแนกประเภทข้อมูลสูงขึ้น

1.3 ขอบเขตการวิจัย

วิทยานิพนธ์นี้มีขอบเขตของการวิจัย ดังนี้

1.3.1 งานวิจัยนี้พัฒนาวิธีการเพิ่มค่าความแม่นยำในการจำแนกข้อมูลโดยการคัดเลือกกลุ่มลักษณะที่ดีที่สุดจากลักษณะที่เป็นไปได้ทั้งหมด โดยขั้นตอนการวิธีเจเนติกค้นหาคำตอบ และใช้ค่าอัตราส่วนเกนเฉลี่ยเป็นฟังก์ชันประเมินความเหมาะสม แล้วเปรียบเทียบกับประสิทธิภาพความแม่นยำในการจำแนกประเภทแบบใช้ลักษณะทั้งหมด

1.3.2 ชุดข้อมูลสำหรับการทดสอบจาก UCI Machine Learning Repository Data Sets

1.3.3 วิเคราะห์ความสามารถในการจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

1.4 ส่วนประกอบของวิทยานิพนธ์

ส่วนประกอบของวิทยานิพนธ์ฉบับนี้ประกอบด้วย

บทที่ 2 กล่าวถึง ทฤษฎีและงานวิจัยที่เกี่ยวข้อง ของการค้นหาลักษณะ และการประเมินลักษณะความรู้ทั่วไปของการจำแนกประเภทโดยใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

บทที่ 3 อธิบายเกี่ยวกับวิธีการคัดเลือกลักษณะโดยวิธีเจเนติก

บทที่ 4 ทำการศึกษาการทดลองและวิธีการวัดประสิทธภาพของคำตอบที่ได้จากวิธีการที่พัฒนาโดยเปรียบเทียบวิธีการจำแนกประเภทโดยลักษณะทั้งหมด กับลักษณะที่คัดเลือกโดยวิธีของงานวิจัย

บทที่ 5 สรุปผลการทดลองและข้อเสนอแนะเกี่ยวกับวิธีการคัดเลือกกลุ่มลักษณะที่ดีที่สุดโดยใช้ขั้นตอนวิธีการทางเจเนติก โดยใช้ค่าอัตราส่วนเกนเฉลี่ยเป็นฟังก์ชันประเมินความเหมาะสม

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องของ การค้นหากลุ่มลักษณะ (Generation Procedures) การประเมินกลุ่มลักษณะ (Evaluation Criterion) และความรู้ทั่วไปของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Back – propagation neural network) ดังต่อไปนี้

2.1 การค้นหากลุ่มลักษณะ

ลักษณะของข้อมูล คือ คุณสมบัติที่ใช้ในการระบุรายละเอียดหรือองค์ประกอบชุดข้อมูลซึ่งแตกต่างไปในแต่ละชุดข้อมูล ลักษณะของข้อมูลที่ดีต้องมีความถูกต้องเชื่อถือได้ หากมีความคลาดเคลื่อนของข้อมูลควรที่จะอยู่ในระดับที่น้อยที่สุดมีความครบถ้วนสมบูรณ์และกะทัดรัด อย่างไรก็ตามการเก็บข้อมูลที่มากเกินความจำเป็นหรือซ้ำซ้อน นอกจากทำให้สิ้นเปลืองพื้นที่หน่วยความจำในการเก็บข้อมูลแล้วยังทำให้สิ้นเปลืองทรัพยากรเวลาและประสิทธิภาพของการนำไปใช้งานหรือการประมวลผลลดลงได้

ด้านศาสตร์ของปัญญาประดิษฐ์มีเทคนิคการค้นหาคำตอบหรือกลุ่มลักษณะเพื่อหากลุ่มลักษณะย่อยที่ดีที่สุดหรือใกล้เคียงคำตอบที่ดี มาเป็นตัวแทนของชุดข้อมูลที่มีขนาดใหญ่ โดยลดลักษณะที่ไม่สำคัญและซ้ำซ้อนออกไป ซึ่งทำให้การทำงานด้านเหมืองข้อมูลไม่ว่าจะเป็นการวิเคราะห์ข้อมูลการเก็บข้อมูลมีประสิทธิภาพมากขึ้น ดังนี้

2.1.1 การค้นหาแบบบอด (Blind search)

2.1.1.1 การค้นหาแบบทั้งหมด (Exhaustive search) ใช้หลักการค้นหาทุกกลุ่มลักษณะที่เป็นไปได้ [1] หรือค้นหาทั้งหมดของมิติข้อมูลภายใต้เกณฑ์การประเมินที่ใช้แล้วเปรียบเทียบว่าคำตอบใดเหมาะสมที่สุด

2.1.1.2 การค้นหาบางส่วน(Partial search) ทำการค้นหาคำตอบเพียงบางส่วนของมิติข้อมูล ซึ่งปัญหาส่วนใหญ่ในทางปัญญาประดิษฐ์มิติข้อมูลจะมีขนาดใหญ่จึงเป็นไปได้ยากในการค้นหาได้ครบทั้งหมด ดังนั้นจึงใช้การค้นหาแค่บางส่วน ดังนั้นจึงมีโอกาสของคำตอบที่ได้อาจจะไม่ใช่คำตอบที่ดีที่สุด เช่น การค้นหาแบบกว้างก่อน (Breadth first search) และ การค้นหาแบบลึกก่อน(Depth first search)

2.1.1.2.1 การค้นหาแบบกว้างก่อน (Breadth first search) [2] เมื่อพิจารณาตามโครงสร้างต้นไม้ (Binary tree) การค้นหาคำตอบจะเข้าถึงแต่ละโหนดตามแนวนอนก่อนโดยเริ่มจากโหนดรากแล้วผ่านไปยังลำดับถัดไปจากซ้ายไปขวา ทำเช่นนี้ไปเรื่อยๆจนกว่าจะพบโหนดสุดท้าย ข้อดีของการค้นหาคำตอบโดยวิธีนี้คือ รับประกันได้ว่าพบคำตอบที่แน่นอนและได้ทำตอบที่ดีที่สุด ไม่ติดในเส้นทางที่ลึกมาก ส่วนข้อเสียคือ หากคำตอบอยู่ในแนวลึกจะสิ้นเปลืองเวลามากในการหาคำตอบ

2.1.1.2.2 การค้นหาแบบลึกก่อน(Depth first search)ทำการค้นหาคำตอบโดยเริ่มที่โหนดรากแล้วผ่านโหนดลูกในด้านใดด้านหนึ่งก่อนจากนั้นผ่านไปยังโหนดลูกของโหนดดังกล่าวอีกครั้งข้อดีของการค้นหาคำตอบโดยวิธีนี้คือใช้ทรัพยากรน้อยกว่า วิธีการค้นหาแบบลึก ถ้าคำตอบอยู่ในระดับลึกจะพบคำตอบโดยไม่ต้องค้นหามากเกินไป ข้อเสียคืออาจติดในเส้นทางที่ลึกมากๆทำให้เสียเวลาในการหาคำตอบที่อยู่ด้านบน

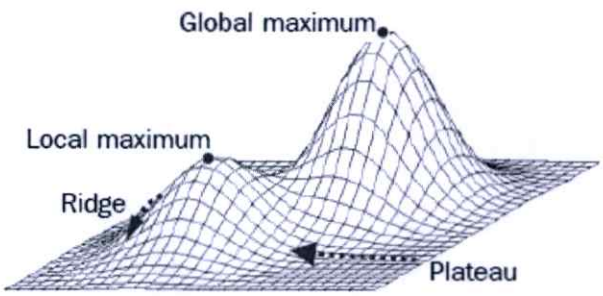
2.1.2 การค้นหาแบบฮิวริสติก (Heuristic Search)

เป็นเทคนิคการนำความรู้ที่เกี่ยวกับปัญหามาช่วยเลือกเส้นทางในการหาคำตอบเหมาะกับปัญหาที่มีจำนวนข้อมูลขนาดใหญ่มากหรือการแก้ปัญหาเอ็นพีบริบูรณ์ทั้งหลายโดยจะพิจารณาจากบางเส้นทางที่น่าจะนำไปสู่คำตอบได้ซึ่งจะลดเวลาในการค้นหา ซึ่งจะทำการวิเคราะห์ปัญหาอย่างมีระเบียบและขั้นตอนการค้นหาด้วยวิธีนี้จะถูกทำซ้ำๆ จนกระทั่งพบคำตอบที่น่าพอใจโดยอาศัยฟังก์ชันฮิวริสติกในการประเมินคำตอบ โดยถ้าฟังก์ฮิวริสติกที่ดีจะทำให้พบคำตอบได้รวดเร็ว ซึ่งอาจเป็นคำตอบที่ดีที่สุดหรือคำตอบที่ใกล้เคียงคำตอบที่ดีที่สุด ทางตรงกันข้ามหากว่าฟังก์ฮิวริสติกไม่ดีอาจทำให้ได้คำตอบที่ไม่ดีหรืออาจไม่พบคำตอบ

2.1.2.1 การค้นหาดีที่สุดก่อน(Best-first search) เป็นการรวมการค้นหาแบบลึกและการค้นหาแบบกว้างมารวมกันเป็นวิธีเดียวซึ่งแต่ละขั้นตอนการค้นหาจะทำการเลือกลักษณะที่ดีที่สุดออกมาก่อนโดยอาศัยฮิวริสติกฟังก์ชันทำหน้าที่เป็นตัววัดผล ทั้งนี้จะเก็บประวัติเส้นทางที่เคยได้ทำการหาคำตอบที่ผ่านมาจึงทำให้ไม่พลาดเส้นทางที่นำไปสู่คำตอบ

2.1.2.2 การค้นหาแบบฮิลไคล์มบิง(Hill climbing) [3] วิธีนี้นำเสนอโดย Pearl ในปี 1984เป็นวิธีการค้นหาข้อมูลที่มีลักษณะคล้ายกับการปีนเขาที่จะต้องเลือกเส้นทางที่จะไปถึงยอดเขาโดยมองก่อนว่ายอดเขาอยู่ที่ใดแล้วพยายามหาเส้นทางที่ดีที่สุดไปยังจุดนั้น โดยหากพบว่าเส้นทางที่ดีกว่าจะเปลี่ยนเลือกเส้นทางนั้นทันทีโดยไม่สนใจเส้นทางปัจจุบัน ซึ่งปัญหาของฮิลไคล์มบิงคือ (Local maximum หรือ foot hill) อาจจะพบตอบแต่อาจไม่ใช่คำตอบที่ดีที่สุด กรณีพบที่ราบสูง (Plateau) การประเมินผลโหนดต่างๆจะเท่ากับโหนดตัวเอง กรณีพบสันเขา (Ridge) คือเมื่อฟังก์ชันฮิวริสติกพาไปพบเส้นทางของการแก้ปัญหาแต่พอสร้างเส้นทางต่อเรื่อยๆกลับพบว่าไม่มีคำตอบที่ดีกว่าและยังไม่พบ

คำตอบเป้าหมาย ซึ่งการแก้ปัญหาของฮิลไคล์มบิงคือ ย้อนกลับไปตั้งต้นเลือกเส้นทางใหม่โดยไม่เลือกซ้ำทางที่เคยเดินมาแล้ว

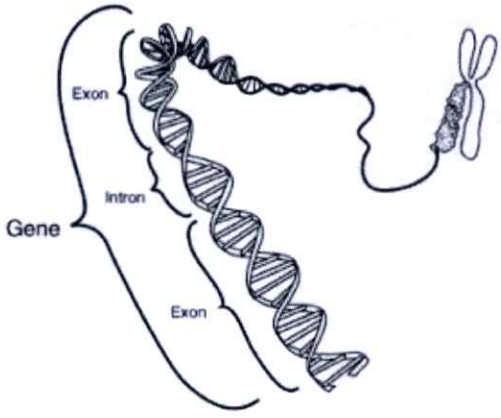


รูปที่ 2.1 ปัญหาของฮิลไคลม์บิง
(ที่มา : ทัศนวรรณ ศูนย์กลาง)

2.1.2.3 การค้นหาแบบการจำลองการอบเหนียว (Simulated annealing) พัฒนาโดย Kirkpatrick,Gelett และ Vecchi ในปี 1983 และ Cerny ในปี 1985[4] เพื่อแก้ปัญหา Local optimum คือ ค่าที่ดีที่สุดเฉพาะที่ซึ่งอาจไม่ใช่ค่าที่ดีสุดของปัญหานั้น โดยใช้แนวคิดของหลักพลศาสตร์ของกระบวนการอบเหนียวในทางฟิสิกส์ ซึ่งเป็นขั้นตอนการหลอมละลายโลหะ โดยการลดอุณหภูมิลงอย่างช้าระหว่างการหลอมเพื่อให้ได้โลหะที่อยู่ในสภาวะที่เหมาะสมที่สุด โดยการจำลองการอบเหนียวจะทำการหาค่าที่ต่ำสุดในลักษณะของยอดเขาหัวกลับ โดยจะยอมให้การค้นหาเดินไปในเส้นทางที่ไม่ดีบ้างในระยะแรก จากนั้นจะค้นหาอย่างละเอียดเมื่อเวลาผ่านไป ซึ่งคำตอบที่ได้จากวิธีนี้จะไม่ขึ้นกับจุดเริ่มต้น

2.1.2.4 การค้นหาแบบเจเนติก เป็นเทคนิคการค้นหาคำตอบที่เหมาะสมกับปัญหา ถูกค้นพบโดย John Holland ในปี 1975 [5][6] โดยอาศัยหลักการจากทฤษฎีทางพันธุศาสตร์การคัดเลือกตามธรรมชาติและวิวัฒนาการของสิ่งมีชีวิตที่จะปรับเปลี่ยนตัวเองเพื่อความอยู่รอดของเผ่าพันธุ์ โดยกระบวนการถ่ายทอดลักษณะทางพันธุกรรมจากรุ่นบรรพบุรุษสู่รุ่นลูกหลานโดยโครโมโซมมีบทบาทสำคัญในการถ่ายทอดลักษณะทางพันธุกรรม ทำให้เกิดการเปลี่ยนแปลงในแต่ละรุ่นที่เรียกว่าวิวัฒนาการ Evolution นั่นคือ Selection, Crossover, Mutation และคำนวณหาค่าความเหมาะสมโดย Fitness Function เพื่อให้เป็นไปตามวัตถุประสงค์ของปัญหาที่กำหนดให้กับโครโมโซมแต่ละตัว

ประโยชน์ของเทคนิคการค้นหาคำตอบนี้ได้พัฒนามาใช้กับงานที่หลากหลาย อาทิเช่น การออกแบบสินค้า การจัดตารางสายการบิน การวางระบบท่อจ่ายน้ำประปา การตรวจจับสแปมเมลล์ การจัดเก็บและการสืบค้นสารสนเทศ เป็นต้น



รูปที่ 2.2 ลักษณะทางพันธุกรรม

(ที่มา : David L. Nelson 2005)

เจเนติกอัลกอริทึมมีองค์ประกอบหลักๆที่สำคัญ 5 องค์ประกอบ[7][8]คือ รูปแบบโครโมโซม (Chromosome Encoding) การสร้างประชากรเริ่มต้น (Initial Population) ฟังก์ชันสำหรับประเมินค่าความเหมาะสม (Fitness Function) การปรับเปลี่ยนองค์ประกอบ (Genetic Operator) และปัจจัยที่ส่งผลต่อการทำงานของเจเนติก (Parameter) ดังต่อไปนี้

รูปแบบโครโมโซม คือรูปแบบที่ใช้ในการนำเสนอทางเลือกที่เป็นไปได้ของแต่ละปัญหา โดยทั่วไปการถอดรหัสนั้นจะขึ้นอยู่กับความเหมาะสมของปัญหานั้นๆ ซึ่งทำให้รูปแบบของโครโมโซมมีความแตกต่างกันออกไปตามแต่ละปัญหา อาทิเช่น

แบบ Binary ทุกตำแหน่งของยีนบนโครโมโซมจะมีค่าเป็นบิต 0 หรือ 1

โครโมโซม A : 00110011101110	โครโมโซม B : 10111011101110

แบบ Direct โครโมโซมจะมีค่าบางค่า ที่สามารถเชื่อมโยงกับปัญหาได้ เช่น ตัวอักษร จำนวนจริง หรือ คำสั่ง

โครโมโซม A : 1.93 0.45 1.67 2.44	โครโมโซม B : back right left back

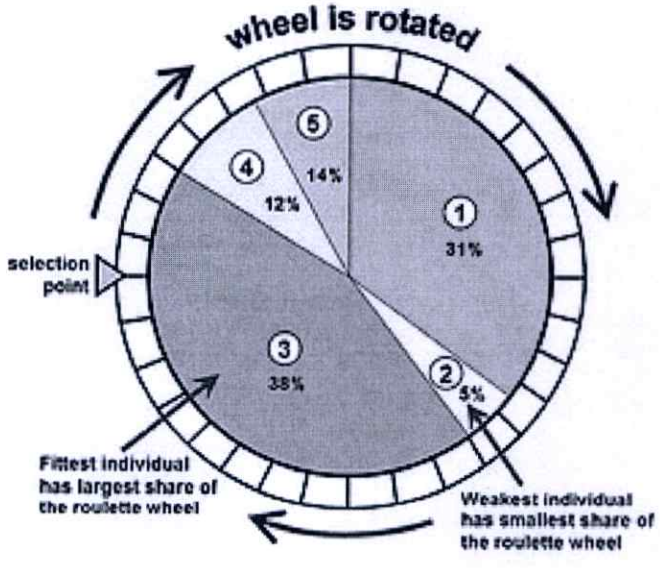
สร้างประชากรเริ่มต้น เพื่อที่จะนำเข้าสู่จุดเริ่มต้นของวิวัฒนาการซึ่งเป็นขั้นตอนแรกของกระบวนการถ่ายทอดลักษณะทางพันธุกรรม โดยส่วนใหญ่จะสร้างจากวิธีการสุ่ม

ฟังก์ชันสำหรับประเมินค่าความเหมาะสม เพื่อให้คะแนนแต่ละคำตอบโครโมโซมทุกตัวจะมีค่าความเหมาะสมของตัวเองเพื่อใช้สำหรับพิจารณาความเหมาะสมในการคัดเลือกนำมาใช้ในการสืบทอดพันธุกรรมสำหรับสร้างประชากรรุ่นใหม่ ซึ่งสมการที่ใช้วัดจะเหมาะสมกับแต่ละปัญหา

การปรับเปลี่ยนองค์ประกอบ ใช้ในการการพัฒนาไปสู่คำตอบที่ดีขึ้นโดยกระบวนการการคัดเลือก (Selection) การไขว้เปลี่ยน (Crossover) และการกลายพันธุ์ (Mutation)

การคัดเลือก การคัดเลือกโครโมโซมเพื่อให้เกิดการอยู่รอดของสิ่งมีชีวิตตามทฤษฎีของ Charles Darwin โดยมีวิธีการคัดเลือกหลายวิธีเช่น แบบ Roulette Wheel การคัดเลือกแบบ Ranking และ การคัดเลือกแบบ Tournament เป็นต้น

- เทคนิค Roulette Wheel Selection เทคนิคนี้นำมาเปรียบเทียบได้กับวงล้อ Roulette ที่มีช่องสล็อตขนาดไม่เท่ากัน โดยช่องสล็อตที่มีขนาดใหญ่จะเทียบได้กับโครโมโซมที่มีค่าความเหมาะสมมาก สามารถที่จะมีโอกาสถูกคัดเลือกไปเป็นประชากรรุ่นต่อไปได้มากขณะที่ช่องสล็อตที่มีขนาดเล็กโอกาสของโครโมโซมที่มีค่าความเหมาะสมน้อย จะถูกคัดเลือกไปสู่ประชากรรุ่นต่อไปได้น้อย ซึ่งสล็อตแต่ละช่องได้จากค่าอัตราส่วนความเหมาะสมของโครโมโซมแต่ละตัว กับค่าความเหมาะสมรวมของโครโมโซม ทุกตัว การคัดเลือกโครโมโซม จะทำการกำหนดจุดคงที่ไว้ 1 จุดแล้วทำการหมุนวงล้อ Roulette เมื่อวงล้อหยุดที่จุดใด โครโมโซมที่ถูกแทนในสล็อต ช่องนั้นจะถูกคัดเลือกไปเป็นประชากรรุ่นต่อไป ดังรูป 2.3



รูปที่ 2.3 การทำงานของวงล้อ Roulette Wheel Selection

(ที่มา : John Dalton 2012)

- เทคนิค Rank Based Selection เทคนิคนี้จะทำการเลือกโครโมโซมโดยไม่สนใจค่าความเหมาะสมของโครโมโซม จะสนใจเพียงลำดับของโครโมโซม วิธีนี้จะทำให้โครโมโซมที่มีค่าความเหมาะสมน้อย มีโอกาสถูกคัดเลือกมากขึ้น

- เทคนิค Tournament Selection เทคนิคนี้จะทำการคัดเลือกประชากรโดยเลียนแบบการแข่งขัน วิธีการจะเริ่มจากการสุ่มโครโมโซมมาสองตัวหรือจำนวนที่ต้องการ จากนั้นหาค่าความน่าจะเป็นโดยการสุ่มตัวเลข ถ้าค่าตัวเลขมีค่ามากกว่าเลขที่กำหนด จะคัดเลือกโครโมโซมที่มีค่าความเหมาะสมมาก ในทางกลับกันถ้าสุ่มเลขได้ค่าน้อยกว่าค่าตัวเลขที่กำหนด จะคัดเลือกโครโมโซมที่มีค่าความเหมาะสมน้อย ไปสร้างเป็นประชากรรุ่นต่อไป

การไขว้เปลี่ยน เป็นการสร้างโครโมโซมใหม่ จากโครโมโซมพ่อแม่จำนวน 2 ตัวแล้วทำการสุ่มคัดเลือกตำแหน่ง ทั้งของพ่อและแม่ แล้วทำการสลับข้อมูลในโครโมโซมของพ่อและแม่ เพื่อสร้างโครโมโซมลูกใหม่ 2 ตัว ดังนี้

- การสลับที่แบบ One-Point Crossover

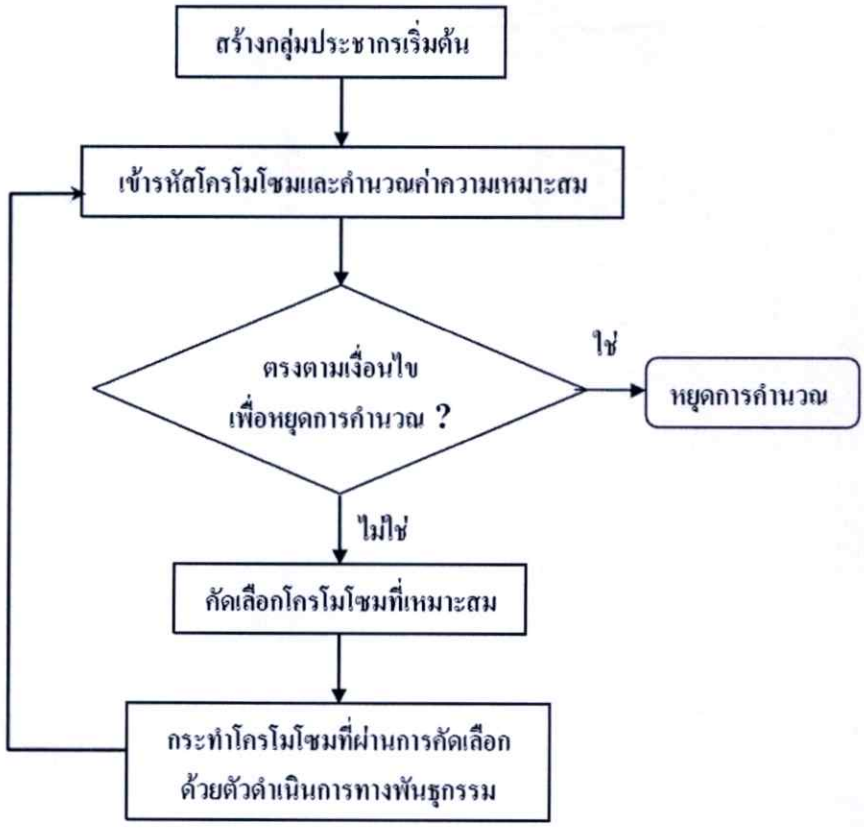
- การสลับที่แบบ Two-Point Crossover

- การสลับที่แบบ Uniform Crossover การสลับตำแหน่งโดยเทคนิคนี้ทุกตำแหน่งบิตของโครโมโซมจะมีโอกาสได้สลับทั้งหมด โดยอาศัยหลักความน่าจะเป็นในการสุ่มเลือกสลับตำแหน่ง

การกลายพันธุ์ เป็นการสร้างโครโมโซมใหม่จากโครโมโซมเดิมโดยจะเปลี่ยนแปลงบิตในโครโมโซมเพียงเล็กน้อย สำหรับเจเนติกการกลายพันธุ์โอกาสเกิดค่อนข้างน้อยส่วนใหญ่จะกำหนดความน่าจะเป็นให้เกิดการกลายพันธุ์ระหว่าง 0% - 1% เพื่อคงไว้ซึ่งการเรียนรู้เลียนแบบ

ธรรมชาติของสิ่งมีชีวิต หากการกลายพันธุ์เกิดขึ้นในทุกตำแหน่งโครโมโซมหรือเกิดขึ้นในทุกรุ่นของประชากรเป็นแบบ 100% วิธีการค้นหาแบบเจเนติกก็จะไม่แตกต่างกับวิธีการค้นหาแบบสุ่ม ตัวอย่างการกลายพันธุ์ต่อตำแหน่งโครโมโซม

ปัจจัยที่ส่งผลต่อการทำงานของเจเนติก เช่น ขนาดของประชากรหากมีจำนวนมากกว่าเกินไปจะทำให้สิ้นเปลืองเวลาในการประมวลผล แต่หากน้อยเกินไปอาจทำให้ลู่เข้าสู่คำตอบที่เป็น Global Minimum ได้ช้าเกินไป ฟังก์ชันสำหรับประเมินความเหมาะสมของคำตอบ ส่วนนี้สำคัญอย่างยิ่งต่อกระบวนการหาคำตอบโดยวิธีเจเนติก ฟังก์ชันที่เหมาะสมกับปัญหาจะทำให้ค้นหาคำตอบที่ดีที่สุดหรือคำตอบที่ใกล้เคียงคำตอบที่ดีได้ซึ่งทำให้วิธีเจเนติกมีประสิทธิภาพมากขึ้น หากว่าฟังก์ชันไม่สอดคล้องกับปัญหาดังกล่าวอาจทำให้ไม่พบคำตอบหรือได้คำตอบที่แย่ นอกจากนี้ยังมี ความน่าจะเป็นของการไขว้เปลี่ยน หรือ ความน่าจะเป็นของการกลายพันธุ์ และ จำนวนรุ่นของประชากร ทั้งนี้ขึ้นอยู่กับความเหมาะสมของแต่ละปัญหา



รูปที่ 2.4 ขั้นตอนวิธีการเจเนติก

2.2 การประเมินลักษณะ

การประเมินความสำคัญของลักษณะ หมายถึง กระบวนการตรวจสอบหรือพิจารณาตัดสินลักษณะเพื่อกำหนดความถูกต้องหรือความเหมาะสม โดยมีเทคนิคที่นิยมใช้ดังนี้

2.2.1 Distance Measures

เป็นการประเมินความสำคัญของลักษณะจากระยะห่าง เช่น Relief algorithm พัฒนาโดยKira&Rendell ในปี 1992 [9] โดยหลักการคือสุ่มออบเจ็คขึ้นมาหนึ่งตัว จากนั้นทำการหาออบเจ็คที่ใกล้เคียงที่สุดในคลาสเดียวกันและต่างคลาสกัน โดยใช้ Euclid distance เพื่อนำออบเจ็คดังกล่าวมาคำนวณเป็นค่าความสำคัญของแต่ละลักษณะซึ่งมีค่าเท่ากับผลต่างระหว่างความน่าจะเป็นของผลต่างระหว่างค่าลักษณะของออบเจ็คที่สุ่มได้ กับ ออบเจ็คที่ใกล้เคียงที่สุดที่ต่างคลาสกันกับความน่าจะเป็นของผลต่างระหว่างค่าลักษณะของออบเจ็คที่สุ่มได้กับออบเจ็คที่ใกล้เคียงสุดที่อยู่ในคลาสเดียวกันดังสมการที่ 2.1

$$w_f = P(\text{different value of } f|\text{different class}) - P(\text{different value of } f|\text{same class}) \quad (2.1)$$

แต่มีข้อเสียคือใช้ได้กับชุดข้อมูลที่แบ่งเป็น 2 คลาสเท่านั้น ในปี 1994 Kononenko มีการพัฒนาเป็น Relief-F เพื่อเพิ่มประสิทธิภาพและสามารถใช้ได้กับข้อมูลที่เป็น Multi Class

2.2.2 Dependence Measures

การประเมินลักษณะจากการวัดสหสัมพันธ์ [10] หรือการวัดที่คล้ายคลึงกันของลักษณะเงื่อนไขกับลักษณะตัดสินใจ การวัดนี้สามารถทำนายค่าของลักษณะหนึ่งจากลักษณะอื่นที่ใกล้เคียงได้ โดยใช้ค่า correlation coefficient (cc) หากค่า cc สูงจะสามารถจำแนกกลุ่มได้ดี

2.2.3 Consistency Measures

การประเมินลักษณะจากการวัดความน่าเชื่อถือของกลุ่มข้อมูล และการใช้ Min-Features bias ในการเลือกลักษณะย่อย การวัดนี้พยายามหาจำนวนลักษณะที่น้อยที่สุดซึ่งเป็นตัวแบ่งแยกกลุ่มที่มีความสอดคล้องจากลักษณะทั้งหมด โดยความสอดคล้องนั้นถูกกำหนดโดย 2 instance ที่มีลักษณะเหมือนกันแต่ต่างกลุ่มกัน

2.2.4 Information Measures

ประเมินลักษณะจากความรู้ทฤษฎีสารสนเทศ ซึ่งมีการคำนวณไม่ซับซ้อนโดยการใช้หลักการสร้างต้นไม้ตัดสินใจ สิ่งที่ควรพิจารณาคือส่วนของลักษณะที่จะทำเป็นโหนดรากของต้นไม้ ซึ่งชี้วัดความสามารถในการจำแนกกลุ่มได้ดีมากน้อยเพียงใด เช่น

2.2.4.1 Gini Index [11] การประเมินค่าที่บ่งบอกว่าลักษณะใดเหมาะสมเป็นลักษณะโดยวัดจากค่าความไม่บริสุทธิ์ ในแต่ละลักษณะ แล้วทำการเปรียบเทียบกับลักษณะอื่นๆเพื่อหาลักษณะที่มีค่า Gini ที่น้อยที่สุดเป็นลักษณะสำคัญ

2.2.4.2 Information Gain [12][13]ใช้จากความรู้ทฤษฎีสารสนเทศ จะพิจารณาจากค่าความน่าจะเป็นของแต่ละลักษณะที่เป็นไปได้แล้ววัดค่าความไร้ระเบียบ (Entropy) เพื่อคัดเลือกถ้าลักษณะใดให้ค่าเกนที่สูงสุด แสดงว่าลักษณะนั้นสามารถจำแนกกลุ่มได้ดีที่สุดซึ่ง Information Gain มีการคำนวณที่ไม่ซับซ้อนแต่ยังมีความเอนเอียงในการประเมิน

2.2.4.3 Gain Ratio Criterion [14] ค่าอัตราส่วนเกนเป็นวิธีประเมินลักษณะที่ใช้วิธีเช่นเดียวกับ Information Gain แต่เพิ่มการหาค่าสารสนเทศการแบ่งแยก (Split Information) เพื่อแก้ไขความเอนเอียงซึ่งจะมีความละเอียดมากขึ้นโดยวิธีการคำนวณค่ามาตรฐานอัตราส่วนเกนดังนี้

1. วัดค่าความไร้ระเบียบของข้อมูลดังสมการ

$$E(S) = -\sum_{c=1}^{N} p(S_C) \times log_2\, p(S_C) \tag{2.2}$$

2. วัดค่า Information Gain เพื่อสร้างลำดับ ดังสมการ

$$Gain(S,V) = E(S) - \sum_{v \epsilon value(v)} \frac{|S_v|}{S} \times E(S_V) \tag{2.3}$$

3. ลดค่าความเอนเอียง ดังสมการ

$$SplitInfo(S,V) = \sum_{i=1}^{m} -\frac{|S_i|}{S} \times log_2 \frac{|S_i|}{S} \tag{2.4}$$

4. หาค่า Gain Ratio Criterion ด้วยค่าการ Split Info ดังสมการ

$$GainRatio(S,V) = \frac{Gain(S,V)}{SplitInfo(S,V)} \tag{2.5}$$

เมื่อ S คือ ลักษณะตัดสินใจ และ V คือ ลักษณะเงื่อนไข

2.2.5 Correlation Based Feature Selection

การประเมินลักษณะจากการจัดอันดับกลุ่มลักษณะย่อยของขนาดของข้อมูล [15] ซึ่งกลุ่มลักษณะย่อยของข้อมูลจะมีความสัมพันธ์กันสูงกับคลาสเดียวกัน และจะไม่มีความสัมพันธ์กับคลาสอื่นๆ ลักษณะของข้อมูลที่ไม่เกี่ยวข้องกันจะถูกคัดออก ซึ่งทำให้ข้อมูลที่ซ้ำซ้อนถูกขจัดออกไปดังสมการที่ 2.6

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \tag{2.6}$$

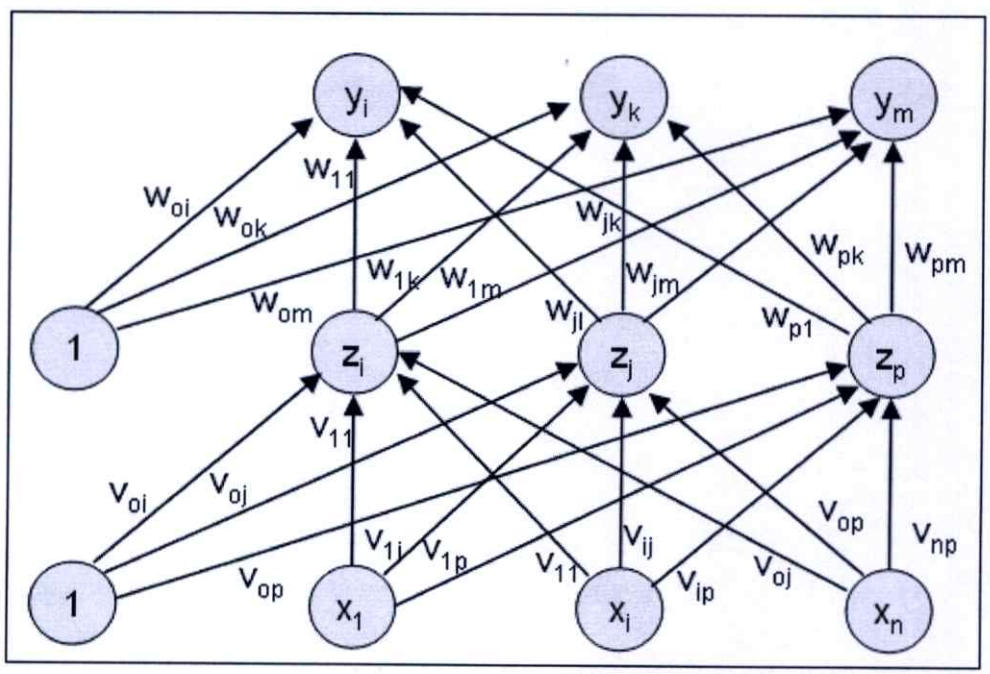
เมื่อ M_s คือ ค่าที่ค้นหาได้ของลักษณะข้อมูลกลุ่มย่อย S ที่ประกอบด้วยลักษณะข้อมูล k

$\overline{r_{cf}}$ คือ ค่าเฉลี่ยความสัมพันธ์ของตัวแปรกับคลาส ($f \epsilon S$)

$\overline{r_{ff}}$ คือ ค่าเฉลี่ยความสัมพันธ์ระหว่างลักษณะของข้อมูล

2.3 โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ [16] เป็นวิธีการเรียนรู้แบบโครงข่ายหลายชั้น (Multilayer neural network) ประกอบไปด้วย ชั้นนำเข้า (Input layer) ชั้นซ่อน (Hidden layer) และชั้นนำออก (Output layer) ดังรูปที่ 2.5 ซึ่งจะมีชุดข้อมูลสำหรับสอน (Training set) และชุดข้อมูลทดสอบ (Test set) วิธีการเรียนรู้แบบแพร่ย้อนกลับนี้จะมีฟังก์ชันความผิดพลาด (Error Function) ในการปรับค่าน้ำหนักของแต่ละชั้นในการเรียนรู้ โดยทุกโหนดของชั้นข้อมูลจะมีเส้นเชื่อมของค่าน้ำหนักเพื่อส่งสัญญาณไปยังโหนดแต่ละชั้นโดยจะมีกระบวนการส่งค่าย้อนกลับสำหรับการสอน ประกอบด้วย 2 ส่วน คือ การส่งผ่านไปข้างหน้า (Forward Pass) และการส่งผ่านย้อนกลับ (Backward Pass) การส่งผ่านไปข้างหน้า ข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นนำเข้าและจะส่งผ่านจากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นนำออกและการส่งผ่านแบบย้อนกลับจะทำการส่งกลับค่าน้ำหนักในทิศทางตรงกันข้ามเพื่อกลับไปแก้ไขค่าความผิดพลาดให้ลดลงถึงระดับที่น่าพอใจโดยแสดงวิธีการเรียนรู้ของโครงข่ายมีดังต่อไปนี้



รูปที่ 2.5 วิธีการเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

(ที่มา : S N Sivanandam 2006)

ค่าตัวแปรที่ใช้ในการเรียนรู้

x คือ ค่าข้อมูลนำเข้า ($x_1, \ldots, x_i, \ldots, x_n$)

t คือ ค่าผลลัพธ์เป้าหมาย ($t_1, \ldots, t_k, \ldots, t_m$)

δ_k คือ ค่าความผิดพลาดชั้นข้อมูลนำออก y_k

δ_j คือ ค่าความผิดพลาดชั้นซ่อน z_j

α คือ อัตราการเรียนรู้

V_{oj} คือ ค่าความลำเอียงในชั้นซ่อน j

z_j คือ ชั้นซ่อน j

w_{ok} คือ ค่าความลำเอียงในชั้นนำข้อมูลออก k

y_k คือ ชั้นนำข้อมูลออก

วิธีการเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับประกอบด้วย 4 ขั้นตอนหลัก ดังต่อไปนี้

1) ขั้นตอนการกำหนดค่าเริ่มต้น
2) ขั้นตอนการส่งผ่านไปข้างหน้า (Feed forward)
3) ขั้นตอนการส่งผ่านค่าความผิดพลาดย้อนกลับ (Back Propagation of errors)
4) ขั้นตอนการปรับค่าน้ำหนัก (Weight) และค่าความลำเอียง (Biases)

ในขั้นตอนแรก จะทำการกำหนดค่า Weight เริ่มต้นโดยการสุ่มค่าตัวเลขที่มีค่าต่ำๆ และในขั้นตอนของการส่งผ่านไปข้างหน้าจะทำการรับเข้ามูลเข้า (X_i) เข้ามายังโครงข่ายและส่งผ่านไปยังแต่ละโหนดในชั้นซ่อน $Z_1 \ldots Z_p$ โดยแต่ละโหนดจะทำการคำนวณผลรวมค่าน้ำหนักและข้อมูลที่รับเข้ามา แล้วส่งผ่านข้อมูลไปยังชั้นนำข้อมูลออกโดยจะผ่านฟังก์ชันกระตุ้น (Sigmoid Function) เพื่อให้ค่าคำตอบของผลลัพธ์ที่เหมาะสม ในขั้นตอนการส่งผ่านค่าความผิดพลาดย้อนกลับ จะทำการหาค่าความผิดพลาดของผลลัพธ์ เพื่อให้ตรงกับค่าเป้าหมายที่ต้องการ(Target Value t_k) เพื่อใช้ในการปรับค่าน้ำหนักโดยมีการคำนวณสองส่วน คือชั้นข้อมูลนำเข้าไปยังชั้นซ่อนและชั้นซ่อนไปยังชั้นนำข้อมูลออก จากนั้นเข้าสู่ขั้นตอนการปรับค่าน้ำหนักและค่าความลำเอียงเพื่อให้ได้ค่าน้ำหนักที่เหมาะสมดังกระบวนการทำงานต่อไปนี้

ขั้นตอนการกำหนดค่าเริ่มต้นของค่าน้ำหนัก

ขั้นตอนที่ 1 กำหนดค่าเริ่มต้นของน้ำหนักด้วยการสุ่มค่าตัวเลขต่ำๆ

ขั้นตอนการส่งผ่านไปข้างหน้า (Feed forward)

ขั้นตอนที่ 2 รับข้อมูลนำเข้าแล้วส่งผ่านไปยังแต่ละโหนดในชั้นซ่อน

ขั้นตอนที่ 3 คำนวณค่าน้ำหนักของข้อมูลนำเข้าในชั้นซ่อน ดังสมการ 2.7

$$Z_{-inj} = V_{oj} + \sum_{i=1}^{n} x_i \, v_{ij} \tag{2.7}$$

แล้วผ่านฟังก์ชันกระตุ้นดังสมการ 2.8

$$Z_j = f(Z_{inj}) \tag{2.8}$$

แล้วส่งผ่านข้อมูลไปยังชั้นนำข้อมูลออก

ขั้นตอนที่ 4 คำนวณค่าน้ำหนักในชั้นนำข้อมูลออก ดังสมการ 2.9

$$y_{-ink} = W_{ok} + \sum_{i=1}^{p} Z_j \, W_{jk} \tag{2.9}$$

แล้วผ่านฟังก์ชันกระตุ้นดังสมการ 2.10

$$Y_k = f(y_{-ink}) \tag{2.10}$$

ขั้นตอนการส่งผ่านค่าความผิดพลาดย้อนกลับ (Back Propagation of errors)

ขั้นตอนที่ 5 รับค่าเป้าหมาย (Target pattern) เข้ามาในแต่ละข้อมูลนำเข้า เพื่อคำนวณหาค่าความผิดพลาด ดังสมการ 2.11

$$\delta_k = (t_k - y_k) f(y_{-ink}) \tag{2.11}$$

ขั้นตอนที่ 6 หาค่าความผิดพลาดระหว่างชั้นนำเข้าไปยังชั้นซ่อนและชั้นซ่อนไปยังชั้นนำข้อมูลออก ดังสมการ 2.12

$$\delta_{-inj} = \sum_{k=1}^{m} \delta_j \, W_{ik} \tag{2.12}$$

และคำนวณช่วงความผิดพลาด ดังสมการ 2.13

$$\delta_i = \delta_{-inj} f(Z_{-inj}) \tag{2.13}$$

ขั้นตอนการปรับค่าน้ำหนัก (Weight) และค่าความลำเอียง (Biases)

ขั้นตอนที่ 7 คำนวณหาค่าน้ำหนักในชั้นนำข้อมูลออกและค่าความลำเอียงแล้วทำการปรับค่าน้ำหนักและค่าความลำเอียง

คำนวณหาค่าน้ำหนัก ดังสมการ

$$\Delta W_{ik} = \alpha \delta_k Z_j \tag{2.14}$$

ค่าความลำเอียง ดังสมการ

$$\Delta W_{ok} = \alpha \delta_k \tag{2.15}$$

ดังนั้นจะได้

$$W_{jk}(new) = W_{jk}(old) + \Delta W_{ik}, W_{ok}(new) = W_{ok}(old) + \Delta W_{ok} \tag{2.16}$$

คำนวณหาค่าน้ำหนักชั้นซ่อนและค่าความลำเอียงแล้วทำการปรับค่าน้ำหนักและค่าความลำเอียง

คำนวณหาค่าน้ำหนัก ดังสมการ 2.17

$$\Delta V_{ij} = \alpha \delta_j x_i \tag{2.17}$$

ค่าความลำเอียง ดังสมการ 2.18

$$\Delta V_{oj} = \alpha \delta_j \tag{2.18}$$

ดังนั้นจะได้

$$V_{ij}(new) = V_{ij}(old) + \Delta V_{ij}, V_{oj}(new) = V_{oj}(old) + \Delta V_{oj} \tag{2.19}$$

ขั้นตอนที่ 8 การหยุดเงื่อนไขเมื่อความผิดพลาดน้อยที่สุด หรือตรงตามเงื่อนไขที่กำหนดไว้

เมื่อฝึกสอนเรียบร้อยแล้วจากนั้นนำคำตอบที่ได้ไปคำนวณโดยฟังก์ชั่นการแปลงข้อมูล (Threshold Function) เพื่อแปลงค่าไปเป็นคำตอบ 1 และ 0 ดังสมการตัวอย่าง

$$f(x) = \begin{cases} 1 & if\ x \geq T \\ 0 & if\ x < T \end{cases} \quad (2.20)$$

โดย T คือค่ากำหนดช่วงผลลัพธ์ ถ้าค่าผลลัพธ์ที่ได้น้อยกว่าค่า Threshold คำตอบเป็น 0 แต่ถ้าค่าผลลัพธ์ มากกว่าหรือเท่ากับค่า Threshold คำตอบเป็น 1

2.4 งานวิจัยที่เกี่ยวข้อง

การคัดเลือกลักษณะมีแนวคิดและงานวิจัยที่หลากหลายและน่าสนใจซึ่งงานวิจัยที่เกี่ยวข้องกับการคัดเลือกลักษณะจากที่ได้ศึกษามีดังต่อไปนี้

1. Ahmed Al-Ani [17] นำเสนอการคัดเลือกลักษณะโดยวิธี Ant Colony Optimization พบว่าวิธีที่นำเสนอให้ผลความแม่นยำที่ 84.22% ซึ่งสูงกว่าวิธี Genetic Algorithm ที่ 83.49%

2. Hongtro Zhang [18] นำเสนอการคัดเลือกลักษณะของแมลงในที่เก็บเมล็ดพันธุ์โดยวิธี Particle Swarm Optimization ร่วมกับการจำแนกประเภทด้วยโครงข่ายประสาทเทียม Support Vector Machines พบว่าให้ความแม่นยำในการเรียนรู้ถึง 95.5% และสามารถคัดเลือกลักษณะที่สำคัญจากทั้งหมด 17 ลักษณะเหลือเพียง 7 ลักษณะ

3. Reza Azm [19] นำเสนอวิธีการคัดเลือกลักษณะด้วยวิธี Hybrid Genetic Algorithm ร่วมกับ Simulated annealing สำหรับการจำแนกภาษาเปอร์เซียจากลายมือพบว่าวิธีดังกล่าวสามารถจำแนกภาพที่มีความซับซ้อนและอัตราการเรียนรู้ที่ดีขึ้น

4. Jianwen Xie [20] นำเสนอการคัดเลือกลักษณะโดยกฎความสัมพันธ์ด้วยวิธีApriori Algorithm พบว่า การคัดเลือกลักษณะโดยกฎความสัมพันธ์ให้ผลความถูกต้องในระดับที่ยอมรับได้แต่ยังคงมีความซับซ้อนของอัลกอริทึมค่อนข้างสูง

5. นรินทร์ พนาวาส [21] ได้ศึกษาการจำแนกมะเร็งเม็ดเลือดขาวโดยใช้เทคนิคการลดมิติข้อมูลข้อมูลด้วย Chi-square ผลจากการทดลองพบว่าการลดมิติข้อมูลด้วยวิธี Chi-square ให้ผลความถูกต้องที่ดีเมื่อทดสอบความถูกต้องในการเรียนรู้ของโครงข่ายประสาทเทียม Support Vector Machine และK-Nearest Neighbor ที่จำนวน 300 มิติ ค่าความถูกต้องเท่ากันที่ 98.61% การเรียนรู้ของโครงข่ายประสาทเทียม Naïve - Bayes ที่จำนวน 4000 มิติ ค่าความถูกต้อง 98.61% และ Dicision Tree เท่ากับ 93.06%

6. ภัทรพงศ์ พงศ์ภัทรกานต์ [22] นำเสนอการเปรียบประสิทธิภาพในการจำแนกข้อมูลแบบจำลอง C5.0,CART,SVM และ SVM ร่วมกับ C5.0 พบว่าแบบจำลอง SVM ผสมผสานกับ C5.0 มี

ความสามารถในการจำแนกประเภทข้อมูลประเภทไบนารีคลาสได้สูงกว่าแบบจำลอง C5.0,CART,SVM เพียงอย่างเดียว

7. สุคนธ์ทิพย์ วงศ์พันธ์ [23] นำเสนอการการเปรียบเทียบการคัดเลือกลักษณะที่เหมาะสมและอัลกอริทึมเพื่อจำแนกพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา โดยเปรียบเทียบเทคนิคการคัดเลือกลักษณะ 3 วิธี ได้แก่ 1. Correlation-based Feature Selection 2. Consistency-based Subset Evaluation และ 3. Wrapper Subset Evaluation เพื่อทำการคัดเลือกลักษณะที่เหมาะสมร่วมกับตัวจำแนกประเภทแบบ Naïve - Bayes เปรียบเทียบกับแบบ Baysian belief network พบว่า การคัดเลือกลักษณะโดยใช้ Wrapper Subset Evaluation ร่วมกับ Baysian belief networkให้ความถูกต้องสูงถึง 82.42%

8. จิราพร สุดใหญ่ [24] นำเสนอการเปรียบเทียบประสิทธิภาพการคัดเลือกลักษณะข้อมูลสำหรับปัญหาการจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบ Extreme Learning Machineโดยขั้นตอนวิธีที่นำมาใช้ลดมิติข้อมูลคือ Principle Component Analysis และ Linear Discriminant Analysis การลดมิติข้อมูลพิจารณาจากค่าไอเกนของข้อมูล พบว่า ชุดข้อมูลที่ผ่านการลดตัวแปรด้วยขั้นตอนวิธี Linear Discriminant Analysis แล้วนำเข้าโครงข่ายประสาทเทียมแบบ Extreme Learning Machineมีความถูกต้องดีกว่าขั้นตอนวิธี Principle Component Analysis

9. ภัทราวุฒิ แสงศิริ[25] นำเสนอการเปรียบเทียบประสิทธิภาพการลดตัวแปรข้อมูลเข้าที่เหมาะสม สำหรับโครงข่ายประสาทเทียมระหว่างเทคนิคการเลือกตัวแปรแบบ Backward Stepwise Feature Selection และ Principle Component Analysis เพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็ง โดยนำผลลัพธ์ที่ได้มาทดสอบกับโครงข่ายประสาทเทียม พบว่าเทคนิคการเลือกตัวแปรแบบแบบ Backward Stepwise Feature Selection ให้ความแม่นยำที่สูงกว่าเทคนิค Principle Component Analysis

10. ภรัณยา อำมฤครัตน์[26] นำเสนอการเปรียบเทียบการลดมิติข้อมูลด้วยวิธี Principle Component Analysis และ Correlation-based Feature Selection ร่วมกับวิธีการจำแนกข้อมูลแบบโครงข่ายประสาทเทียมแบบ Multi-Layer Perceptron เปรียบเทียบ Support Vector Machine พบว่าการลดมิติข้อมูลแบบ Correlation-based Feature Selection ร่วมกับวิธีการจำแนกข้อมูลแบบ Multi-Layer Perceptron มีประสิทธิภาพดีกว่าโมเดลแบบอื่นๆ

11. ฐิติมา เกษมศรีธนาวัฒน์ [27] นำเสนอการจำแนกความคิดเห็นโดยใช้ตัวจำแนกแบบ Naïve - Bayes ร่วมกับการคัดเลือกลักษณะด้วย Relief algorithm พบว่าความแม่นยำของการเรียนรู้บนทัศนคติที่เกี่ยวข้องกับหนังสือโดย Naïve Bayes ให้ความแม่นยำที่ดี เมื่อใช้ร่วมกับการคัดเลือกลักษณะโดย Relief

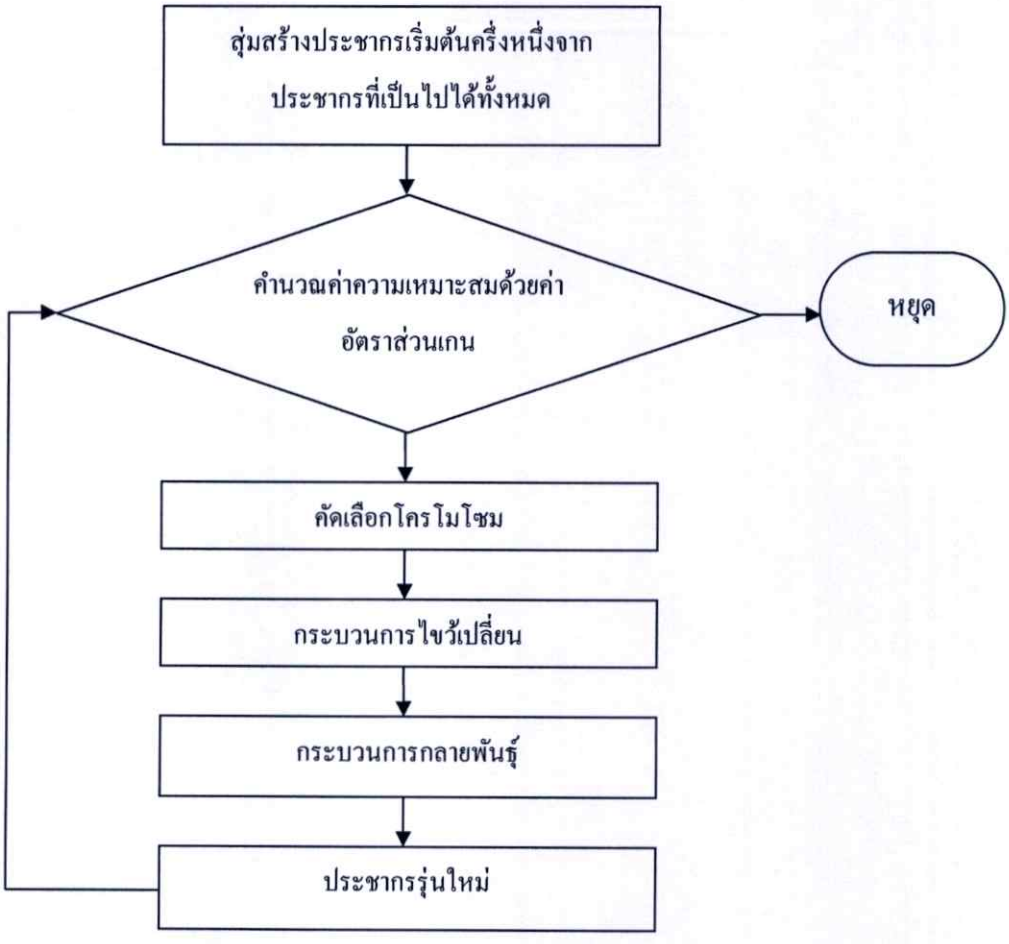
บทที่ 3

การคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนโดยใช้ขั้นตอนวิธีเจเนติก

ในบทที่ 2 กล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องของการค้นหาลักษณะ การประเมินลักษณะ และความรู้ทั่วไปของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับแล้ว ในบทนี้จะกล่าวถึงการคัดเลือกกลุ่มลักษณะด้วยค่าอัตราส่วนเกน โดยใช้ขั้นตอนวิธีเจเนติกและการวัดประสิทธิภาพความแม่นยำของกลุ่มลักษณะที่คัดเลือกได้

3.1 การคัดเลือกลักษณะ

การคัดเลือกลักษณะเพื่อลดข้อมูลไม่ตรงประเด็น และซับซ้อน ที่ส่งผลต่อประสิทธิภาพของการวิเคราะห์ข้อมูล โดยทั่วไปการประเมินกลุ่มย่อยเพื่อที่จะวัดความสำคัญของลักษณะจะต้องทำทุกกลุ่มย่อยที่เป็นได้ทั้งหมด ซึ่งในความเป็นจริงหากจำนวนข้อมูลมีขนาดมหาศาลย่อมทำได้ยากที่จะหาความสำคัญของทุกย่อยที่เป็นไปได้ทั้งหมด งานวิจัยนี้จึงใช้ขั้นตอนวิธีเจเนติกในการค้นหากลุ่มลักษณะดังกล่าว เพื่อที่จะลดการหาทุกกลุ่มลักษณะย่อยที่เป็นไปได้ทั้งหมด แล้วได้กลุ่มลักษณะย่อยที่ยอมรับได้และเพิ่มความแม่นยำให้กับการจำแนกประเภทโดยการคัดเลือกกลุ่มลักษณะด้วยขั้นตอนวิธีเจเนติกดังต่อไปนี้



รูปที่ 3.1 ขั้นตอนการเลือกกลุ่มลักษณะโดยขั้นตอนวิธีเจเนติก

1. สร้างประชากรเริ่มต้นโดยวิธีการสุ่มเลือกประชากรครึ่งหนึ่งจากประชากรที่เป็นไปได้ทั้งหมด
2. คำนวณหาฟังชันก์ความเหมาะสมของประชากรเริ่มต้นแต่ละตัวจากนั้นทำการวัดการกระจายตัวของฟังชันก์ความเหมาะสมด้วยส่วนเบี่ยงเบนมาตรฐาน (standard deviation)
3. ถ้าการกระจายตัวของฟังชันก์ความเหมาะสมของประชากรแต่ละตัวมากกว่า 0.0001 จะเข้าสู่กระบวนการการไขว้เปลี่ยนและกระบวนการกลายพันธุ์เพื่อสร้างเป็นประชากรรุ่นต่อไป แต่ถ้าการกระจายตัวของฟังชันก์ความเหมาะสมน้อยกว่า 0.0001 แสดงว่ากระบวนการค้นหาลู่เข้าสู่คำตอบจะทำการหยุดหรือจะทำการหยุดเมื่อผลิตประชากรครบรุ่นที่กำหนด
4. ค่าพารามิเตอร์ในการปรับเปลี่ยนสำหรับการไขว้เปลี่ยน 0.2 และ การกลายพันธุ์ 0.01 ขนาดประชากรครึ่งหนึ่งของประชากรทั้งหมด

3.2 ตัวอย่างการคำนวณ

ตารางที่ 3.1 ตารางการตัดสินใจของชุดข้อมูลอากาศ

No.	Outlook	Temperature	Humidity	Windy	Play
1.	Sunny	85.0	85.0	False	No
2.	Sunny	80.0	90.0	True	No
3.	Overcast	83.0	86.0	False	Yes
4.	Rainy	70.0	96.0	False	Yes
5.	Rainy	68.0	80.0	False	Yes
6.	Rainy	65.0	70.0	True	No
7.	Overcast	64.0	65.0	True	Yes
8.	Sunny	72.0	95.0	False	No
9.	Sunny	69.0	70.0	False	Yes
10.	Rainy	75.0	80.0	False	Yes
11.	Sunny	75.0	70.0	True	Yes
12.	Overcast	72.0	90.0	True	Yes
13.	Overcast	81.0	75.0	False	Yes
14.	Rainy	71.0	91.0	True	No

ตัวอย่างที่ 3.1 การค้นหากลุ่มลักษณะโดยขั้นตอนวิธีเจเนติก

สมมุติลักษณะเงื่อนไขที่เป็นไปได้ในชุดข้อมูลWeather เท่ากับ $2^n - 1$ ซึ่งมีลักษณะเงื่อนไข 4 ลักษณะV={outlook, temperature, humidity, windy} ดังนั้นโอกาสที่เป็นไปได้ของประชากรที่เกิดขึ้นดังตารางที่ 3.2

ตารางที่ 3.2 โอกาสของประชากรทั้งหมดที่เป็นไปได้

1000	0001	1001	0011	1011
0100	1100	0110	1110	0111
0010	1010	0101	1101	1111

ทำสร้างประชากรโดยวิธีสุ่มเพียง $\frac{N}{2} - 1$ ของประชากรทั้งหมดดังนั้นโอกาสการเกิดประชากรเริ่มต้นจะได้ 6 ตัวโดยไม่คำนึงถึงโอกาสการเกิดซ้ำดังตารางที่ 3.3

ตารางที่ 3.3 การสร้างประชากร

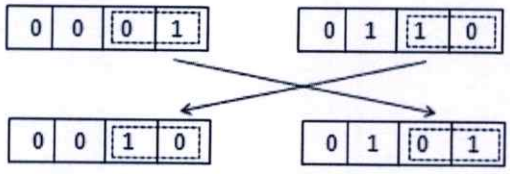
0 0 0 1	1 0 0 1	1 1 0 0
1 1 0 0	0 1 1 0	1 0 1 0

คำนวณความเหมาะสมด้วยฟังก์ชันหาความเหมาะสมเฉลี่ยของประชากรแต่ละตัว ดังตารางที่ 3.4

ตารางที่ 3.4 ความเหมาะสมเฉลี่ยของประชากรแต่ละตัว

0 0 0 1	0.15642756242117517
1 1 0 0	0.269645931096278685
1 0 0 1	0.180851870176935495
0 1 1 0	0.207600838735547015
1 1 0 0	0.269645931096278685
1 0 1 0	0.172596251271031085

ทำการวัดการกระจายตัวของค่าเฉลี่ยความเหมาะสมด้วยส่วนเบี่ยงเบนมาตรฐานถ้าการกระจายของค่าเฉลี่ยความเหมาะสมมากกว่า 0.0001 เข้าสู่กระบวนการไขว้ เปลี่ยนต่อไปโดยการสุ่มจับคู่เลือกประชากร 20% ของจำนวนประชากรที่ถูกสร้าง



รูปที่ 3.2 ไขว้เปลี่ยน

เข้าสู่กระบวนการกลายพันธุ์เมื่อความน่าจะเป็นเท่ากับ 0.01 โดยกำหนดให้กระบวนการกลายพันธุ์จะเลือกสุ่มตำแหน่งกลับบิตในสายโครโมโซมนั้น 1 ตำแหน่งตัวอย่าง ดังรูปที่ 3.3

รูปที่ 3.3 การกลายพันธุ์

ได้ประชากรรุ่นใหม่แล้ววนเข้าสู่ขั้นตอนที่ 2 – 4 จนกว่าจะได้ค่ากระจายของค่าเฉลี่ยความเหมาะสมน้อยกว่า 0.0001 หรือรุ่นประชากรที่กำหนดไว้คือ 500 รุ่น

<u>ตัวอย่างที่ 3.2</u> การคำนวณหาค่าความเหมาะสมด้วยอัตราส่วนเกนของกลุ่มลักษณะ

หลักการหาค่าความเหมาะสมด้วยอัตราส่วนเกนของกลุ่มลักษณะ คือ หากค่าความเหมาะสมของกลุ่มลักษณะดังกล่าวมีค่าสูงกว่ากลุ่มลักษณะอื่นๆ สรุปได้ว่ากลุ่มลักษณะดังกล่าวจะมีความสามารถในการจำแนกกลุ่มได้ดี ดังนี้

1. คำนวณค่าความไร้ระเบียบของข้อมูลดังสมการ

$$E(S) = -\sum_{c=1}^{N} p(S_c) \times \log_2 p(S_c) \tag{3.1}$$

เมื่อ c คือ ขอบเขตความสนใจ S คือ ลักษณะตัดสินใจ และ V คือ ลักษณะเงื่อนไข จากตารางที่ 3.1

ลักษณะตัดสินใจ S = Play [Yes:9,No:5] และ

ลักษณะเงื่อนไข V = {Outlook,Temperature,Humidity,Windy}

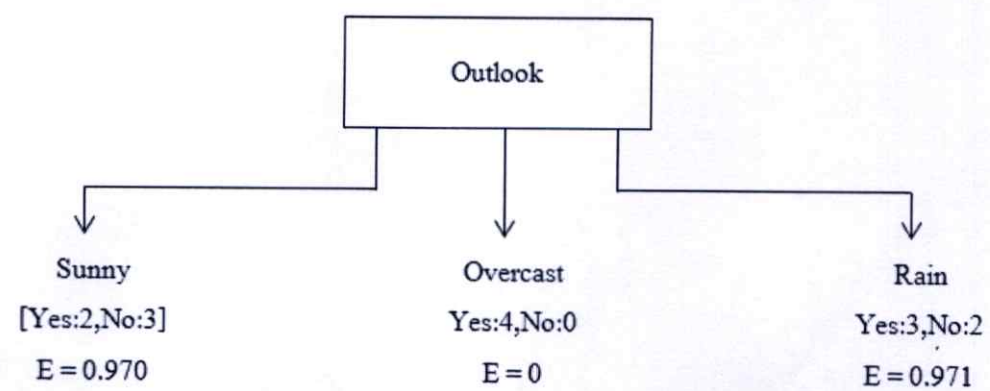
จากสมการสามารถวัดค่าความไร้ระเบียบของข้อมูลได้ดังนี้

$$Entropy(S) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)$$

$$= 0.940$$

2. คำนวณค่า Information Gain ดังสมการ

$$Gain(S,V) = E(S) - \sum_{v \epsilon value(v)} \frac{|S_v|}{S} \times E(S_v) \tag{3.2}$$



$$Gain(S, Outlook) = 0.940 - \left(\frac{5}{14}\right) \times 0.970 - \left(\frac{4}{14}\right) \times 0 - \left(\frac{5}{14}\right) \times 0.971$$

$$= 0.247$$

3. คำนวณค่า $SplitInfo(S, V)$ ดังสมการ

$$SplitInfo(S, V) = \sum_{i=1}^{m} -\frac{|s_i|}{S} \times \log_2 \frac{S_i}{S} \tag{3.3}$$

$$SplitInfo(S, V) = -\left|\frac{5}{14}\right| \log_2 \left|\frac{5}{14}\right| - \left|\frac{4}{14}\right| \log_2 \left|\frac{4}{14}\right| - \left|\frac{5}{14}\right| \log_2 \left|\frac{5}{14}\right|$$

$$= 1.5774$$

4. คำนวณค่า Gain Ratio Criterion ด้วยค่าการ Split Info ดังสมการ

$$GainRatio(S, V) = \frac{Gain(S, V)}{SplitInfo(S, V)} \tag{3.4}$$

$$GainRatio(S, V) = \frac{0.247}{1.5774}$$

$$= 0.1565$$

ดังนั้น จะได้ค่าความสำคัญของลักษณะ Outlook = 0.1565 ซึ่งจะมีกระบวนการคำนวณเช่นนี้จนครบทุกลักษณะเงื่อนไข

สำหรับงานวิจัยนี้ได้ประยุกต์ผลลัพธ์ที่ได้จากสมการที่ 3.4 เป็นฟังก์ชันความเหมาะสมของวิธีเจเนติกเพื่อวัดค่าความเหมาะสมเฉลี่ยของประชากรตัวนั้นๆ หากว่าค่าเฉลี่ยของประชากรตัวดังกล่าวมีค่าสูงกว่าประชากรตัวอื่นๆ แสดงว่ากลุ่มลักษณะที่คัดเลือกนั้นมีความสำคัญ ดังสมการที่ 3.5

$$fn = \frac{\sum GainRatio(S, V)}{V} \tag{3.5}$$

บทที่ 4

ผลการทดลอง

จากการทดลองในบทที่ 3 ได้กล่าวถึงกระบวนการทำงานของการคัดเลือกกลุ่มลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก ในบทนี้กล่าวถึงชุดข้อมูลที่ใช้ทดลอง วิธีการทดลอง การจำแนกประเภทด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ และการวัดความแม่นยำในการจำแนกประเภท

4.1 ชุดข้อมูลที่ใช้ในการทดลอง

การทดลองนี้ได้ใช้ชุดข้อมูลจาก UCI Machine Learning Repository [28] จำนวน 5 ชุดข้อมูล ได้แก่ Heart-Statlog (HS) , Labor , CongressVoting-1984 (CV), Wisconsin-Breast-Cancer (WBC) และ Glass

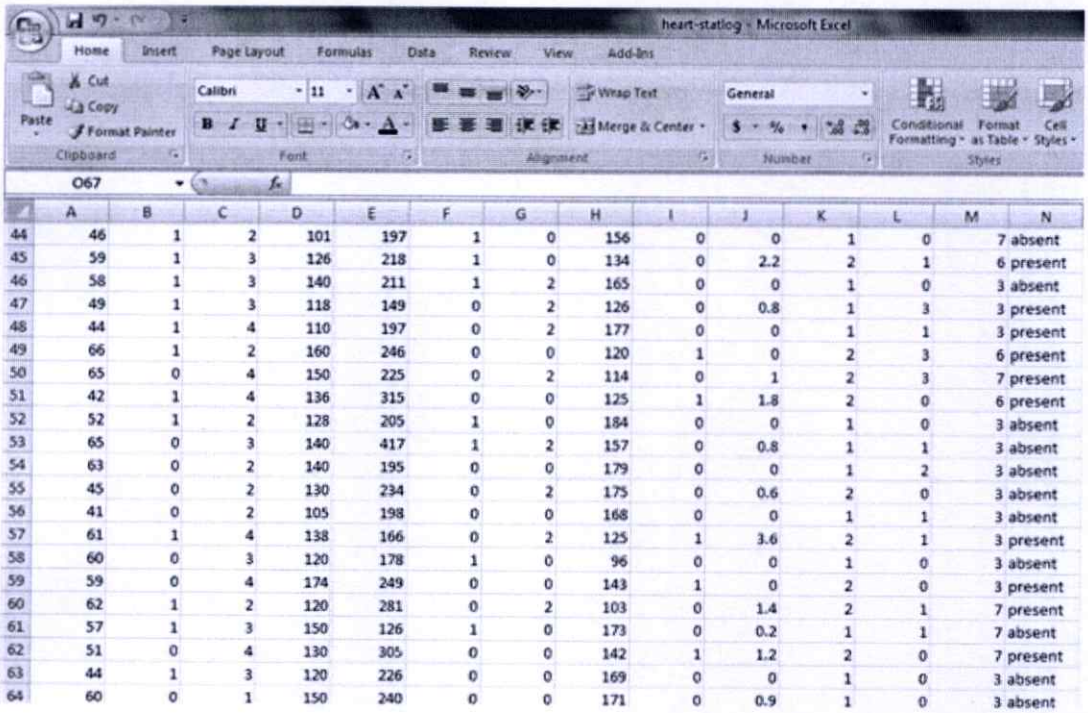
ตารางที่ 4.1 ชุดข้อมูลการทดลอง

Dataset	No. Attribute	Attribute Type	No. Instance
HS	13	Categorical,Real	270
Labor	16	Integer,Real	57
CV	16	Categorical	435
WBC	9	Integer	699
Glass	9	Real	214

4.2 การเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูลจะทำการแปลงข้อมูลออกเป็น 2 นามสกุล คือ นามสกุล .csv เพื่อใช้สำหรับนำชุดข้อมูลเข้าฐานข้อมูล MySQL Database Version 5.0.45 เพื่อที่จะใช้ในการทดลองในส่วนของกลุ่มลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติกโดยจะตัด

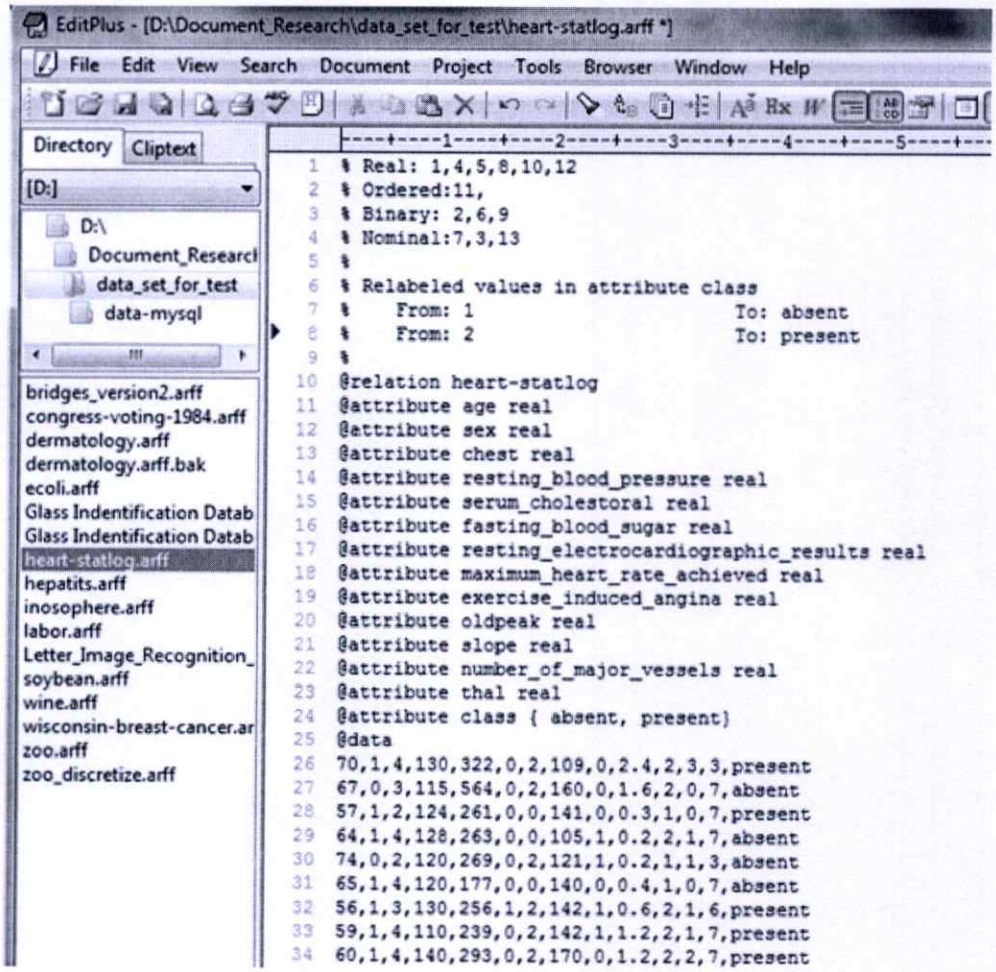
ส่วนของชื่อคอลัมน์ออกและจัดให้ส่วนของคอลัมน์ตัดสินใจอยู่ในคอลัมน์สุดท้ายในทุกชุดข้อมูลโดยโปรแกรมMicrosoft Excel ดังรูปที่ 4.1



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
44	46	1	2	101	197	1	0	156	0	0	1	0	7	absent
45	59	1	3	126	218	1	0	134	0	2.2	2	1	6	present
46	58	1	3	140	211	1	2	165	0	0	1	0	3	absent
47	49	1	3	118	149	0	2	126	0	0.8	1	3	3	present
48	44	1	4	110	197	0	2	177	0	0	1	1	3	present
49	66	1	2	160	246	0	0	120	1	0	2	3	6	present
50	65	0	4	150	225	0	2	114	0	1	2	3	7	present
51	42	1	4	136	315	0	0	125	1	1.8	2	0	6	present
52	52	1	2	128	205	1	0	184	0	0	1	0	3	absent
53	65	0	3	140	417	1	2	157	0	0.8	1	1	3	absent
54	63	0	2	140	195	0	0	179	0	0	1	2	3	absent
55	45	0	2	130	234	0	2	175	0	0.6	2	0	3	absent
56	41	0	2	105	198	0	0	168	0	0	1	1	3	absent
57	61	1	4	138	166	0	2	125	1	3.6	2	1	3	present
58	60	0	3	120	178	1	0	96	0	0	1	0	3	absent
59	59	0	4	174	249	0	0	143	1	0	2	0	3	present
60	62	1	2	120	281	0	2	103	0	1.4	2	1	7	present
61	57	1	3	150	126	1	0	173	0	0.2	1	1	7	absent
62	51	0	4	130	305	0	0	142	1	1.2	2	0	7	present
63	44	1	3	120	226	0	0	169	0	0	1	0	3	absent
64	60	0	1	150	240	0	0	171	0	0.9	1	0	3	absent

รูปที่ 4.1 แปลงข้อมูลให้เป็นนามสกุล .csv

สำหรับการแปลงเป็นนามสกุล .arff เพื่อสำหรับทดสอบความแม่นยำของการจำแนกประเภทของโครงข่ายประสาทเทียมโดยทำการทดสอบกับโปรแกรม Weka ซึ่งต้องจัดข้อมูลโดยโปรแกรม EditPlus ให้อยู่ในรูปแบบที่โปรแกรม Weka อ่านข้อมูลได้โดยทั่วไปใช้สัญลักษณ์ดังนี้ เครื่องหมาย % หมายถึง การเขียนอธิบายส่วนต่างๆของชุดข้อมูลโดยไม่ต้องการให้โปรแกรมอ่าน (comment) ดังบรรทัดที่ 1-9 เครื่องหมาย @ บรรทัดที่ 10 เป็นต้นไปหมายถึงส่วนที่จะให้โปรแกรมอ่านข้อมูลซึ่งจะต้องประกอบไปด้วย @relation ตามด้วยชื่อตาราง @attribute ตามด้วยชื่อลักษณะ @data ตามด้วยข้อมูลในบรรทัดถัดลงมาและใช้สัญลักษณ์จุลภาค (" , ")ในการแบ่งวรรคข้อมูล ดังรูปที่ 4.2



รูปที่ 4.2 แปลงข้อมูลให้เป็นนามสกุล .arff

4.3 การแบ่งข้อมูลสอนและทดสอบ

การแบ่งข้อมูลสอนและทดสอบโดยวิธี k-fold cross validation ซึ่งจะแบ่งข้อมูลออกเป็นกลุ่มจำนวน k กลุ่ม ในตอนแรกจากนั้นเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอนจากนั้นจะสลับข้อมูลกลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่นๆที่เหลือเป็นชุดสอน สลับเช่นนี้ไปเรื่อยๆจนครบ k กลุ่ม หลังจากนั้นในขั้นตอนสุดท้ายจะหาค่าเฉลี่ยของค่าความถูกต้องในแต่ละกลุ่ม ซึ่งวิธีการนี้ข้อมูลทุกตัวจะได้เป็นทั้งชุดทดสอบและชุดสอน ซึ่งในการทดลองได้แบ่งข้อมูลแบบ 10 - fold cross-validation แบ่งชุดข้อมูลสอน และชุดข้อมูลทดสอบ

4.4 การวัดประสิทธิภาพ

งานวิจัยนี้ใช้การทดลองแบบ k - fold cross-validation แบ่งชุดข้อมูลสอน และชุดข้อมูลการทดลองแต่ละรอบจะมีการคำนวณหาความแม่นยำ ดังสมการที่ 4.3.1

$$accuracy_i = \frac{Correctly\ Classified\ Instance}{Total\ Number\ of\ Instance} \times 100 \quad (4.1)$$

และหาค่าเฉลี่ยความแม่นยำของผลลัพธ์ทั้งหมดเพื่อใช้ในการเปรียบเทียบประสิทธิภาพ

$$AVG_{accuracy} = \frac{\sum_{i=1}^{n} accuracy_i}{n} \quad (4.2)$$

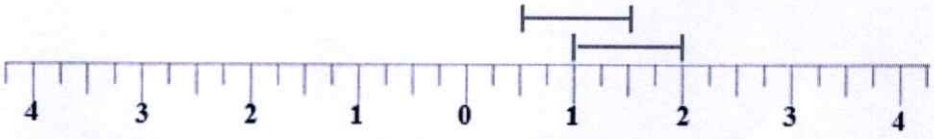
เมื่อ n คือ จำนวน k - fold cross-validation

จากนั้นทำการทดลองสรุปความแม่นยำที่ระดับความเชื่อมั่น (Confidence Level) 95% หรือที่ระดับนัยสำคัญทางสถิติ (Level of Significant) $\alpha = 0.05$ ของค่าเฉลี่ยถ้าความแม่นยำของแต่ละชุดการทดลอง ซึ่งในทางสถิติช่วงระดับความเชื่อมั่นแบ่งออกได้เป็น 5 กรณีดังนี้

ช่วงระดับความเชื่อมั่นคร่อมไปทางขวา (Right overlap) แสดงว่าช่วงระดับความเชื่อมั่นสูงขึ้นอย่างไม่มีนัยสำคัญดังตัวอย่าง[0.5,1.5][1,2] รูปที่ 4.3

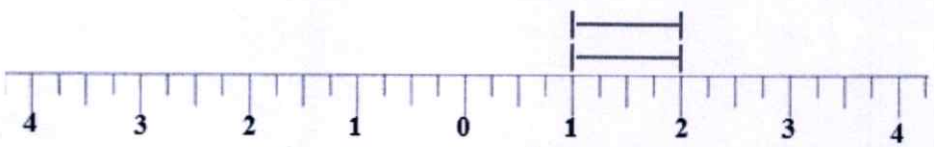
รูปที่ 4.3 ช่วงระดับความเชื่อมั่นคร่อมไปทางขวาอย่างไม่มีนัยสำคัญ

ช่วงระดับความเชื่อมั่นคร่อมไปทางซ้าย (Left overlap) แสดงว่าช่วงระดับความเชื่อมั่นต่ำลงอย่างไม่มีนัยสำคัญดังตัวอย่าง [1,2][0.5,1.5] รูปที่ 4.4



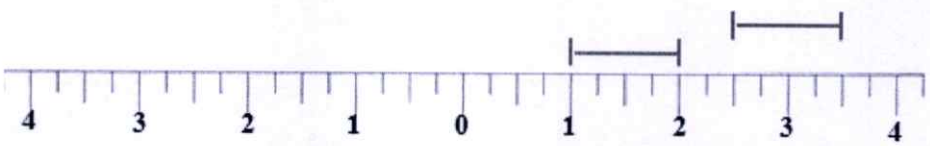
รูปที่ 4.4 ช่วงระดับความเชื่อมั่นคร่อมไปทางซ้ายอย่างไม่มีนัยสำคัญ

ช่วงระดับความเชื่อมั่นเสมอกัน (Balance overlap) แสดงว่าช่วงระดับความเชื่อมั่นเท่ากันดังตัวอย่าง [1,2][1,2] รูปที่ 4.5



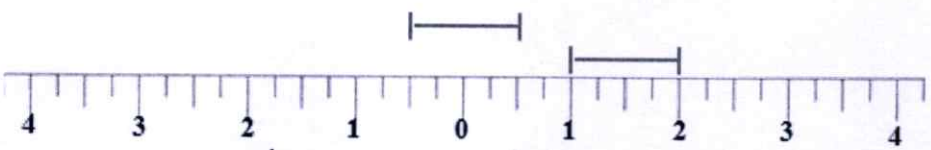
รูปที่ 4.5 ช่วงระดับความเชื่อมั่นเสมอกัน

ช่วงระดับความเชื่อมั่นไม่คร่อมไปทางขวา (Right non overlap) แสดงว่าช่วงระดับความเชื่อมั่นสูงขึ้นอย่างมีนัยสำคัญดังตัวอย่าง [1,2][2.5,3.5] รูปที่ 4.6



รูปที่ 4.6 ช่วงระดับความเชื่อมั่นไม่คร่อมไปทางขวา

ช่วงระดับความเชื่อมั่นไม่คร่อมไปทางซ้าย (Left non overlap) แสดงว่าช่วงระดับความเชื่อมั่นต่ำลงอย่างมีนัยสำคัญดังตัวอย่าง [1,2][-0.5,0.5] รูปที่ 4.7



รูปที่ 4.7 ช่วงระดับความเชื่อมั่นไม่คร่อมไปทางซ้าย

โดยสมการสำหรับหาช่วงระดับความเชื่อมั่นดังนี้

$$\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}} \qquad (4.3)$$

เปิดตาราง Z ที่ค่า Z = 1.96

โดยที่ $\bar{X}$ คือ ค่าเฉลี่ยของความแม่นยำ

S คือ ส่วนเบี่ยงเบนมาตรฐานความผิดพลาด

n คือ จำนวน fold ที่ใช้ในการแบ่งชุดข้อมูลสอนและทดสอบ

4.5 ทดสอบการจำแนกโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

การทดลองนี้ได้นำผลการคัดเลือกกลุ่มลักษณะที่ได้ มาทดสอบกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในโปรแกรม Weka version 3.6.2 ตั้งค่าการทดสอบ โดย ค่าอัตราการเรียนรู้ 0.3 ค่าโมเมนตัม 0.2 ค่าน้ำหนักเริ่มต้น 0 จำนวนรอบการสอน 500 รอบ

4.6 ผลการทดลอง

จำนวนลักษณะที่ได้หลังทำการคัดเลือกลักษณะแล้วเท่ากับ 3,3,7,7,6 ตามลำดับ ซึ่งสามารถลดจำนวนลักษณะได้ถึง 76.93%,81.25%,56.25%,22.22% และ 33.33% ของจำนวนลักษณะทั้งหมดของแต่ละชุดข้อมูล ดังตารางที่ 4.2

ตารางที่ 4.2 จำนวนลักษณะที่ได้จากการคัดเลือกกลุ่มลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก

Dataset	Attribute	Attribute name	Fitness function Avg	(%)Reduct
HS	3	exercise induced angina, number of major vessels, thal	0.14177475816318574	76.93 %
Labor	3	pension, standby-pay, longterm-disability-assistance	0.31298346774067990	81.25%
CV	7	adoption-of-the-budget-resolution, physician-fee-freeze, el-salvador-aid, aid-to-nicaraguan-contras, mx-missile, synfuels-corporation-cutback, superfund-right-to-sue	0.35821576245074505	56.25%
WBC	7	clump_thickness, cell_size_uniformity, cell_shape_uniformity, marginal_adhesion, single_epi_cell_size, bare_nuclei, normal_nucleoli	0.23633944486308917	22.22%
Glass	6	Na, Mg, Al, Si, K,Ca	0.25576693422962943	33.33%

จากนั้นนำชุดลักษณะข้อมูลที่ได้จากการคัดเลือกไปทดสอบความแม่นยำโดยการจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ ดังตารางที่ 4.3

ตารางที่ 4.3 ทดสอบความแม่นยำโดยการจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

Dataset	BP	AvgGainRaitoGA+BP
	Correctly	Correctly
HS	78.15 %	**82.96 %**
Labor	87.72 %	**92.98 %**
CV	94.48 %	**96.32 %**
WBC	95.28 %	**96.85 %**
Glass	67.75 %	**70.09 %**

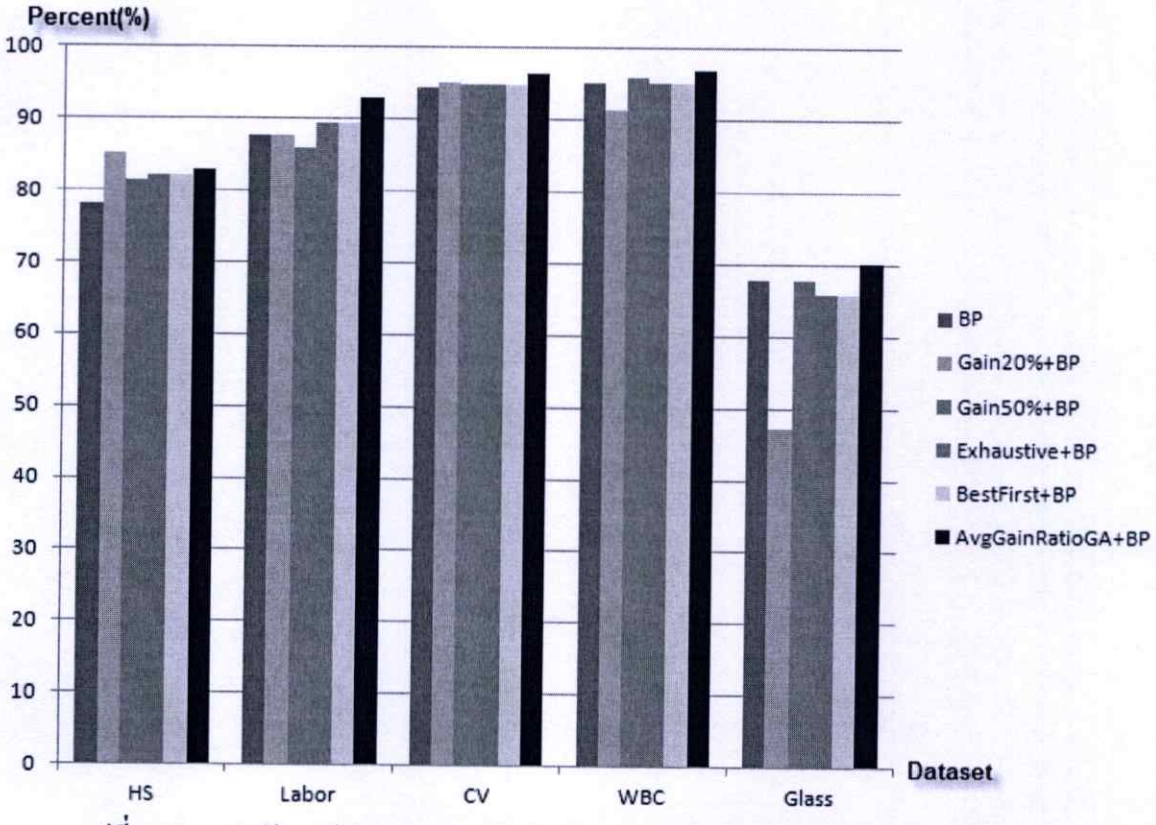
ผลจากการทดสอบความแม่นยำโดยการจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับพบว่าชุดข้อมูล heart-statlog มีความแม่นยำถึง 82.96 % ซึ่งความแม่นยำของข้อมูลก่อนการคัดเลือกลักษณะเพียง 78.15 % เช่นเดียวกันกับชุดข้อมูล Labor ชุดข้อมูล congress-voting-1984 ชุดข้อมูล wisconsin-breast-cancer ชุดข้อมูล Glass ซึ่งมีความแม่นยำในการจำแนกประเภทสูงขึ้นถึง 92.98%, 96.32%, 96.85% และ 70.09 % ตามลำดับ

เมื่อเปรียบเทียบความแม่นยำหลังจากทำการคัดเลือก กับความแม่นยำก่อนทำการคัดเลือกการคัดเลือกลักษณะสำคัญ 20% และ 50% จากลำดับเกนสูงสุดของลักษณะทั้งหมด การคัดเลือกลักษณะด้วยวิธี Exhaustive และวิธี BestFirst ซึ่ง 2 วิธีนี้ได้ใช้เกณฑ์การประเมินความสำคัญ Correlation Based Feature Selection (CfsSubsetEval) ในโปรแกรม Weka สำหรับคัดเลือกลักษณะสำคัญ พบว่าการคัดเลือกกลุ่มลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติกให้ผลความแม่นยำในการจำแนกประเภทสูงกว่าทั้ง 4 วิธี ยกเว้นวิธีการคัดเลือกลักษณะสำคัญจากลำดับเกนสูงสุดที่ 20% ของชุดข้อมูล heart-statlog ที่มีความแม่นยำถึง 85.18 % ดังแสดงตารางที่ 4.4

ตารางที่ 4.4 แสดงการเปรียบเทียบความแม่นยำในการจำแนกประเภท

Dataset	BP	Gain20%+BP	Gain50%+BP	Exhaustive+BP	BestFirst +BP	AvgGainRaitoGA+BP
	Correctly	Correctly	Correctly	Correctly	Correctly	Correctly
HS	78.15 %	**85.18%**	81.48%	82.22 %	82.22 %	**82.96 %**
Labor	87.72 %	87.71%	85.96%	89.47 %	89.47 %	**92.98 %**
CV	94.48 %	95.17%	94.94%	94.94 %	94.94 %	**96.32 %**
WBC	95.28 %	91.41%	95.85%	95.28 %	95.28 %	**96.85 %**
Glass	67.75 %	47.19%	67.75%	65.89 %	65.89 %	**70.09 %**

จากตารางที่ 4.4 แสดงผลการเปรียบเทียบความแม่นยำของโครงข่ายประสาทเทียมโดยการคัดเลือกลักษณะในแต่ละวิธีข้างต้นดังรูปกราฟที่ 4.8



รูปที่ 4.8 กราฟเปรียบเทียบผลความแม่นยำของการจำแนกด้วยโครงข่ายประสาทเทียม

จากการทดลองโดยการคัดเลือกกลุ่มลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติกพบว่าบางชุดข้อมูล เช่น ชุดข้อมูล Heart-Statlog (HS) มีลักษณะที่ให้ค่าความแม่นยำในการเรียนรู้ของโครงข่ายเท่ากันที่ 82.96 % จะพิจารณาจากค่าฟังก์ชันความเหมาะสมที่มากที่สุด เช่นเดียวกันกับกรณีชุดลักษณะที่มีจำนวนลักษณะเท่ากันและค่าความแม่นยำในการเรียนรู้ของโครงข่ายเท่ากันแต่ลักษณะไม่ตรงกันก็จะพิจารณาที่ฟังก์ชันความเหมาะสมก่อนเสมอดังตารางที่ 4.5 จะทำการเลือกลักษณะที่ 1 ซึ่งมีค่าฟังก์ชันความเหมาะสมเฉลี่ยเท่ากับ 0.14177475816318574 มีจำนวน 3 แอตทริบิวต์

ตารางที่ 4.5 การคัดเลือกกลุ่มลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติกของชุดข้อมูล Heart-Statlog (HS)

No.	Attribute	Fitness Function
1.	0000000010011	0.14177475816318574
2.	1010100010111	0.10585499068636509
3.	0010000011011	0.11902937985795307

ส่วนต่อไปนี้แสดงผลการเปรียบเทียบช่วงความแตกต่างของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของการเรียนรู้แบบโครงข่ายประสาทเทียมของชุดข้อมูลทั้ง 5 ชุด ดังตารางที่ 4.6 และตารางถัดไปตามลำดับ ซึ่งชุดข้อมูลทั้ง 5 ชุด มี 4 ชุดข้อมูล ในการทดลองที่ให้ช่วงความเชื่อมั่นที่ระดับนัยสำคัญ 95% ซึ่งสูงกว่าวิธีการอื่นๆอย่างมีนัยสำคัญ โดยมีเพียงหนึ่งชุดการทดลองคือ ชุดข้อมูล heart-statlog ที่เมื่อทำการทดลองโดยวิธีการคัดเลือกลักษณะสำคัญ 20% จากลำดับเกนสูงสุดของลักษณะทั้งหมด นั้นให้ช่วงความแม่นยำที่ระดับนัยสำคัญ 95% สูงกว่าวิธีที่นำเสนอ อย่างไรก็ตามแม้ว่าวิธีที่งานวิจัยนี้นำเสนอให้ผลลัพธ์ที่ไม่ดีเท่าที่ควรกับชุดข้อมูล heart-statlog แต่ข้อดีของวิธีการที่นำเสนอคือ การหาความเหมาะสมของชุดลักษณะ โดยที่ไม่ได้จำกัดหรือระบุขอบเขตของลักษณะอย่างตายตัว ทำให้โอกาสที่จะได้กลุ่มลักษณะที่เหมาะสมเพื่อการจำแนกกลุ่มเพิ่มมากขึ้น

ตารางที่ 4.6 ผลการเปรียบเทียบช่วงความแตกต่างของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Heart-Statlog

Dataset : Heart-Statlog		
Method	**Min**	**Max**
BP	77.88	78.41
Gain20%+BP	**84.96**	**85.4**
Gain50%+BP	81.22	81.73
Exhaustive+BP	81.98	82.47
BestFirst +BP	81.98	82.47
AvgGainRaitoGA+BP	**82.73**	**83.19**

ตารางที่ 4.7 ผลการเปรียบเทียบช่วงความแตกต่างของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Labor

Dataset : Labor		
Method	**Min**	**Max**
BP	87.50	87.93
Gain20%+BP	87.52	87.89
Gain50%+BP	85.75	86.16
Exhaustive+BP	89.28	89.66
BestFirst +BP	89.28	89.66
AvgGainRaitoGA+BP	**92.82**	**93.15**

ตารางที่ 4.8 ผลการเปรียบเทียบช่วงความแตกต่างของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล CongressVoting-1984

Dataset : CongressVoting-1984		
Method	**Min**	**Max**
BP	94.35	94.61
Gain20%+BP	95.04	95.29
Gain50%+BP	94.80	95.07
Exhaustive+BP	94.82	95.06
BestFirst +BP	94.82	95.06
AvgGainRaitoGA+BP	**96.21**	**96.44**

ตารางที่ 4.9 ผลการเปรียบเทียบช่วงความแตกต่างของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Wisconsin-Breast-Cancer

Dataset : Wisconsin-Breast-Cancer		
Method	**Min**	**Max**
BP	95.15	95.4
Gain20%+BP	91.25	91.56
Gain50%+BP	95.73	95.96
Exhaustive+BP	95.15	95.4
BestFirst +BP	95.15	95.4
AvgGainRaitoGA+BP	**96.75**	**96.96**

ตารางที่ 4.10 ผลการเปรียบเทียบช่วงความแตกต่างของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Glass

Dataset : Glass		
Method	**Min**	**Max**
BP	67.59	67.91
Gain20%+BP	47.01	47.36
Gain50%+BP	67.58	67.91
Exhaustive+BP	65.71	66.06
BestFirst +BP	65.71	66.06
AvgGainRaitoGA+BP	**69.93**	**70.25**

บทที่ 5

สรุปและข้อเสนอแนะ

5.1 สรุป

งานวิจัยนี้ทำการคัดเลือกกลุ่มลักษณะที่สำคัญโดยวิธีเจเนติกซึ่งร่วมกับการประเมินลักษณะโดยอัตราส่วนเกนเฉลี่ยเป็นฟังก์ชันความเหมาะสมจากผลการทดลองชุดข้อมูลจำนวน 5 ชุดข้อมูล พบว่า ชุดข้อมูล Labor , CongressVoting-1984, Wisconsin-Breast-Cancer และ Glass ให้ผลความแม่นยำในการจำแนกประเภทข้อมูลสูงขึ้นเมื่อเปรียบเทียบกับการจำแนกประเภทข้อมูลก่อนทำการคัดเลือก การคัดเลือกลักษณะสำคัญร้อยละ20 และร้อยละ 50 จากลำดับเกนสูงสุดของลักษณะทั้งหมด การค้นหาลักษณะด้วยวิธี BestFirst, วิธีExhaustive ซึ่ง 2 วิธีนี้ได้ใช้เกณฑ์การประเมินความสำคัญ Correlation Based Feature Selection (CfsSubsetEval) ในโปรแกรม Weka อย่างมีนัยสำคัญที่ระดับความเชื่อมั่น 95% ส่วนชุดข้อมูล Heart-Statlog ให้ผลความแม่นยำเพียง 82.96 % แม้ว่าจะให้ผลความแม่นสูงกว่า 4 วิธีข้างต้นแต่ต่ำกว่าวิธีการคัดเลือกลักษณะสำคัญ20 % จากลำดับเกนสูงสุดของลักษณะทั้งหมดที่มีความแม่นยำ 85.18 % เมื่อหาช่วงความเชื่อมั่นที่ระดับนัยสำคัญ 95% พบว่าผลความแม่นยำเพียง 82.96 % ไม่มีนัยสำคัญเมื่อเปรียบเทียบกับผลการทดลองของวิธีการคัดเลือกลักษณะสำคัญ 20 % แต่อย่างไรก็ตามการคัดเลือกกลุ่มลักษณะโดยวิธีที่งานวิจัยนำเสนอนั้นหาความเหมาะสมของกลุ่มลักษณะโดยที่ไม่ได้จำกัดหรือระบุขอบเขตที่ตายตัวทำให้โอกาสที่จะได้กลุ่มลักษณะที่เหมาะสมเพื่อการจำแนกข้อมูลเพิ่มมากขึ้น

5.2 ข้อเสนอแนะ

ในการทดลองพบว่าค่าฟังก์ชันความเหมาะสมที่ใช้เป็นเกณฑ์ในการค้นหาและเลือกลักษณะมีผลต่อการคัดเลือกลักษณะที่ดีเนื่องจากว่าค่าฟังก์ชันเป็นตัวแปรสำคัญที่จะทำให้ค้นพบลักษณะที่ดีหรือลักษณะที่ไม่เหมาะสมได้ งานวิจัยครั้งต่อไปมุ่งศึกษา วิธีการประเมินลักษณะวิธีอื่นๆ เช่น Correlation – based Feature Selection (CFS) และวิธีการค้นหาแบบฮิวริสติกแบบอื่นๆ เช่น วิธีการจำลองการอบเหนียว (Simulated Annealing) เพื่อหลีกเลี่ยงจากการติด local หรือวิธีและเปรียบเทียบการจำแนกประเภทข้อมูลด้วยวิธีโครงข่ายประสาทเทียมตัวอื่นๆต่อไป เช่น RBF และ SVM เป็นต้น

เอกสารอ้างอิง

[1] Huan Liu, 2005, "Toward Integrating Feature Selection Algorithm for Classification and Clustering", IEEE Transactions Knowledge And Data Engineering, VOL.17, NO.4.

[2] ขนิษฐา นามี , 2548, "โครงสร้างข้อมูลและอัลกอริทึม Data Structure Algorithm" , 117 .

[3] David J. Stracuzzi,2001,"Computation Methods of Feature Selection ",11-30.

[4] Marc Pirlot, 1996,"General local search methods", European Journal of Operational Research 92,493-511.

[5] Darrell Whitley,1994," A genetic algorithm tutorial ",4,65-85.

[6] D.Beasley,1993,"An Overview of Genetic Algorithms : Part1",University Computing , 15 (2) , 58-69.

[7] Hartmut Pohlheim,2006," Introduction Evolutionary Algorithms: Overview, Methods, and Operators " .

[8] D.Beasley,1993,"An Overview of Genetic Algorithms : Part 2",University Computing,15(4),
170 - 181.

[9] Ron Kohavi,1997, "Wrappers for Feature Subset Selection",AIJ special issue on relevance,7.

[10] M.Dash,H.Liu,1997, "Feature Selection for Classification" Intelligent Data Analysis,131-156.

[11] Laura E,Raileanu ,Kilian Stoffel,2004,"Theoretical Comparison between the Gini Index and Information Gain", Swiss National Science Foundation 41,77–93.

[12] นฤพนต์ ว่องประชานุกูล, 2548, "วิธีที่เหมาะสำหรับการตัด กิ่งต้นไม้ตัดสินใจของการทำเหมืองข้อมูล", วิทยานิพนธ์ปริญญาวิศวกรรมศาสตร์มหาบัณฑิตสาขาวิชาวิศกรรมศาสตร์มหาวิทยาลัยเทคโนโลยีสุรนารี.

เอกสารอ้างอิง(ต่อ)

[13] Matthew N. Anyanwu,“ Comparative Analysis of Serial Decision Tree Classification Algorithms ”, International Journal of Computer Science and Security, Issue (3),230-240.

[14] J.R.QUINLAN, 1986, “Induction of Decision Trees”, Machine Learning 1, 81-106.

[15] Patharawat Saengsiri,2010, “Classification of Leukemia Data Using Ranking and Support Vector Machine” ,KKE Res J (GS) 10 (2) ,10-16.

[16] S N Siyanandam,2006, “Introduction To Neural Networks using Matlab 6.0”,184-189.

[17] Ahmed Al-Ani,2005,“Ant Colony Optimization for Feature Subset Selection”, World Academy of Science, Engineering and Technology 4 , 35-38.

[18] Hongtro Zhang,2009,“Feature Selection for the Stored-grain Insects Based on PSO and SVM”, IEEE Second International Workshop on Knowledge Discovery and Data Mining ,586-589.

[19] Reza Azm,2010,“ A hybrid GA and SA algorithms for feature selection in recognition of hand - printed Farsi characters”, IEEE ,384-387.

[20] Jianwen Xie, 2009,“ Feature Selection Algorithm Based on Association Rules Mining Method” ,IEEE/ACIS International Conference on Computer and Information Science ,357-362.

[21] นรินทร์ พนาวาส,“การจำแนกมะเร็งเม็ดเลือดขาวโดยการใช้เทคนิคการลดมิติข้อมูลด้วย Chi-square”.

[22] ภัทรพงศ์ พงศ์ภัทรกานต์,2009,“การเปรียบเทียบการจำแนกข้อมูลของแบบจำลอง CART,SVM,C5.0 และแบบผสมผสานกัน”,NCCIT The 5th National Conference on Computing and Information Technology,1102-1106.

เอกสารอ้างอิง(ต่อ)

[23] สุคนธ์ทิพย์ วงศ์พันธ์,2551, "การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะเพื่อหาปัจจัยที่มีผลต่อพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา",การประชุมวิชาการเทคโนโลยีและนวัตกรรมสำหรับการพัฒนาอย่างยั่งยืน คณะวิศกรรมศาสตร์ มหาวิทยาลัยขอนแก่น.

[24] จิราพร สุดใหญ่,2554,"การเปรียบเทียบประสิทธิภาพการคัดเลือกคุณลักษณะข้อมูลสำหรับปัญหาการจำแนกประเภทข้อมูลด้วยโครงข่ายประสาทเทียมแบบเอ็กซ์ทรีม",The 4th National Conference on Technical Education,335-340.

[25] ภัทราวุฒิ แสงศิริ,2009,"การเปรียบเทียบประสิทธิภาพการลดตัวแปรข้อมูลเข้าที่เหมาะสมสำหรับโครงข่ายประสาทเทียมระหว่างเทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้นและการวิเคราะห์องค์ประกอบเพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็ง", NCCIT The 5th National Conference on Computing and Information Technology,851-858.

[26] ภรัณยา อำมฤครัตน์,2010,"การเปรียบเทียบประสิทธิภาพการลดมิติข้อมูลและการจำแนกข้อมูลโดยวิธีการทางเครือข่ายประสาทเทียม",การประชุมทางวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาครั้งที่ 11,58-65.

[27] ฐิติมา เกษมศรีธนาวัฒน์,2011,"การจำแนกความคิดเห็นโดยใช้ตัวจำแนกแบบเบย์ร่วมกับการเลือกคุณลักษณะด้วยอัลกอริทึมรีลีฟ",CIT2011&UniNOMS2011.

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

ภาคผนวก ก

งานวิจัยที่ตีพิมพ์

บทคัดย่อ
Abstracts

The 23rd การประชุมวิชาการเสนอผลงานวิจัย ระดับบัณฑิตศึกษาแห่งชาติ ครั้งที่ 23

The 23rd National Graduate Research Conference

คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน

วันที่ 23-24 ธันวาคม 2554

ณ คณะวิทยาศาสตร์และศิลปศาสตร์
มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน

Faculty of Sciences and Liberal Arts
Rajamangala University of Technology Isan

การคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก
A Gain Ratio Based Feature Selection for Classification using Genetic Algorithms

บุญญาพร เข็มปัญญา[1]* วีระ บุญจริง[2]

[1]*สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพ* 10520
[2] *National Centre for Excellence in Mathematics, PERDO, Bangkok 10400, Thailand*
** E-mail: khempanya.b@hotmail.com*

บทคัดย่อ

งานวิจัยนี้เสนอวิธีการคัดเลือกกลุ่มลักษณะโดยพิจารณาค่าอัตราส่วนเกนของแต่ละลักษณะ กลุ่มลักษณะที่ถูกคัดเลือกจะเป็นกลุ่มที่มีค่าอัตราส่วนเกนเฉลี่ยสูงสุด เนื่องจาก การค้นหากลุ่มดังกล่าวต้องทำการประเมินกลุ่มย่อยที่เป็นไปได้ทั้งหมดซึ่งในปัญหาที่มีลักษณะจำนวนมากทำได้ยาก งานวิจัยนี้จึงใช้ขั้นตอนวิธีเจเนติกในการค้นหากลุ่มลักษณะดังกล่าว เมื่อนำกลุ่มลักษณะที่ได้มาทดสอบกับโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับเพื่อจำแนกประเภทข้อมูล พบว่าโครงข่ายนี้ให้ความแม่นยำในการจำแนกประเภทสูงกว่าโครงข่ายที่สร้างจากลักษณะทั้งหมดอย่างมีนัยสำคัญ

คำสำคัญ : การคัดเลือกลักษณะ เจเนติกอัลกอริทึม ค่าอัตราส่วนเกน

บทนำ

ปัญหาการคัดเลือกลักษณะสำหรับการจำแนกประเภทข้อมูลเป็นปัญหาการค้นหากลุ่มลักษณะย่อยที่ดีที่สุดตามเกณฑ์ที่กำหนดโดยเกณฑ์ที่ใช้เป็นเกณฑ์ที่สามารถวัดความสำคัญของแต่ละลักษณะได้ อัตราส่วนเกน (Gain Ratio) จัดเป็นเกณฑ์ที่มีการขจัดความเอนเอียงของลักษณะที่มีค่าแตกย่อยมาก งานวิจัยนี้จึงเลือกใช้เกณฑ์ดังกล่าวในการวัดความสำคัญของแต่ละลักษณะต่องานจำแนกประเภท สำหรับการค้นหากลุ่มลักษณะที่ดีที่สุดตามเกณฑ์นั้นต้องมีการประเมินแต่ละกลุ่มจากกลุ่มย่อยที่เป็นไปได้ทั้งหมด ซึ่งในปัญหาที่มีลักษณะจำนวนมากทำได้ยากเนื่องจากมีกลุ่มย่อยที่ต้องประเมินจำนวนมหาศาลเมื่อเทียบกับจำนวนลักษณะ ดังนั้น การค้นหาคำตอบจึงต้องทำโดยใช้ขั้นตอนวิธีการประมาณ งานวิจัยนี้จึงเสนอใช้ขั้นตอนวิธีการเจเนติกในการค้นหาคำตอบดังกล่าว โดยใช้ค่าอัตราส่วนเกนเฉลี่ยเป็นฟังก์ชันประเมินความเหมาะสม ส่วนที่เหลือของบทความวิจัยนี้ประกอบด้วย ตอนที่ 2 เป็นทฤษฎีและงานวิจัยที่เกี่ยวข้อง ตอนที่ 3 กล่าวถึงขั้นตอนวิธีการที่เสนอในการเลือกลักษณะ ตอนที่ 4 และ 5 เป็นวิธีการทดลองและผลการทดลอง ตามลำดับ ตอนที่ 6 สรุปบทความวิจัยนี้

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การคัดเลือกลักษณะเป็นขั้นตอนหนึ่งที่สำคัญในการเตรียมข้อมูลสำหรับการทำเหมืองข้อมูล (Data Mining) เพื่อลดข้อมูลไม่ตรงประเด็น และซับซ้อน ที่ส่งผลต่อประสิทธิภาพของการวิเคราะห์ข้อมูล โดยทั่วไปการคัดเลือกลักษณะประกอบด้วย การค้นหาลักษณะ (Generation Procedures) และการประเมินลักษณะ (Evaluation Criterion)

การค้นหาลักษณะ

การค้นหาลักษณะเพื่อหากลุ่มลักษณะย่อยที่ดีที่สุด หรือใกล้เคียง ดังนี้

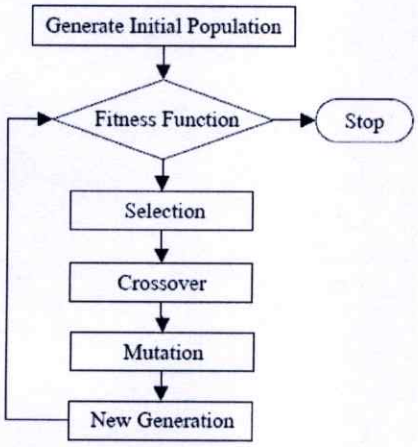
การค้นหาแบบทั้งหมด [1] ใช้หลักการค้นหาทุกลักษณะที่เป็นไปได้ภายใต้เกณฑ์การประเมินที่ใช้ เช่น Exhaustive Search

การค้นหาดีที่สุดก่อน [2] เป็นการรวมการค้นหาแบบลึกและการค้นหาแบบกว้างมารวมกันเป็นวิธีเดียว ซึ่งแต่ละขั้นตอนการค้นหาจะทำการเลือกลักษณะที่ดีที่สุดออกมาก่อน โดยอาศัยฮิวริสติกฟังก์ชัน ทำหน้าที่เป็นตัววัดผล

การค้นหาแบบสุ่ม เป็นวิธีที่นิยมใช้กับปัญหาที่มองหาลักษณะเฉพาะจากโอกาสที่เป็นไปได้ทั้งหมดเพื่อเลือกลักษณะที่ดีที่สุดซึ่งโดยทั่วไปหากต้องการลักษณะที่ดีที่สุดจะต้องทำการเปรียบเทียบกับทุกลักษณะที่เป็นได้จึงมีปัญหาอย่างมากหากข้อมูลมีขนาดใหญ่ ดังนั้นการค้นหาลักษณะโดยวิธีการสุ่มจะช่วยลดการเปรียบเทียบจากทุกลักษณะที่เป็นไปได้ซึ่งอาจได้ลักษณะที่ดีที่สุดหรือคำตอบที่ใกล้เคียงคำตอบที่ดีที่สุด เช่น วิธีเจเนติก

วิธีการเจเนติกเป็นเทคนิคการค้นหาคำตอบที่เหมาะสมโดยอาศัยหลักการจากทฤษฎีวิวัฒนาการของสิ่งมีชีวิตที่จะปรับเปลี่ยนตัวเองเพื่อความอยู่รอด สำหรับองค์ประกอบหลักๆ คือการแปลงโครโมโซม(Chromosome Encoding) โดยจะทำ

การแปลงปัญหาให้อยู่ในรูปแบบโครโมโซมแล้วสร้างประชากรเริ่มต้น(Initial Population) โดยการสุ่มเลือกจากกลุ่มประชากรทั้งหมดที่มีอยู่เพื่อใช้เป็นชุดเริ่มต้นของขั้นตอนขั้นตอนเจเนติกแล้วทำการประเมินค่าความเหมาะสม(Fitness function)เพื่อให้คะแนนสำหรับคำตอบที่เป็นไปได้ของโครโมโซมทุกตัวจากนั้นดำเนินการพัฒนาไปสู่คำตอบที่ดีขึ้นโดยกระบวนการการคัดเลือก(Selection),การไขว้เปลี่ยน(Crossover) และการกลายพันธุ์(Mutation)โดยมีปัจจัยที่ส่งผลต่อการทำงานของเจเนติก(Parameter) เช่น ขนาดของประชากร, ความน่าจะเป็นของการ การไขว้เปลี่ยน หรือความน่าจะเป็นของการกลายพันธุ์เมื่อดำเนินจนครบจะได้ประชากรรุ่นใหม่และนำเข้าสู่ขั้นตอนประเมินค่าความเหมาะสมเช่นเดิมจนได้ค่าที่เหมาะสมหรือสิ้นสุดเงื่อนไขที่กำหนดไว้ ดังรูปที่ 1



รูปที่ 1 ขั้นตอนวิธีการเจเนติก

การประเมินลักษณะ

การประเมินความสำคัญของลักษณะมีเทคนิค ที่นิยมใช้ดังนี้

Distance Measures [3, 4] เป็นการประเมินความสำคัญของลักษณะจากระยะห่าง เช่น Relief algorithm โดยหลักการคือสุ่มออบเจ็คขึ้นมาหนึ่งตัว จากนั้นทำการหาออบเจ็คที่ใกล้เคียงที่สุดในคลาสเดียวกันและต่างคลาสกัน โดยใช้ Euclid distance เพื่อนำออบเจ็คดังกล่าวมาคำนวณเป็นค่าความสำคัญของแต่ละลักษณะซึ่งมีค่าเท่ากับผลต่างระหว่างความน่าจะเป็นของผลต่างระหว่างค่าลักษณะของออบเจ็คที่สุ่มได้ กับ ออบเจ็คที่ใกล้เคียงที่สุดที่ต่างคลาสกัน กับ ความน่าจะเป็นของผลต่างระหว่างค่าลักษณะของออบเจ็คที่สุ่มได้กับออบเจ็คที่ใกล้เคียงสุดที่อยู่ในคลาสเดียวกัน ดังสมการที่ 1

$$w_f = P(\text{different value of } f|\text{different class}) - P(\text{different value of } f|\text{same class}) \quad (1)$$

แต่มีข้อเสียคือใช้ได้กับชุดข้อมูลที่แบ่งเป็น 2 คลาสเท่านั้น ดังนั้นจึงมีการพัฒนาเป็น Relief-F เพื่อเพิ่มประสิทธิภาพและสามารถใช้ได้กับข้อมูลที่เป็น Multi Class

Dependence Measures [5] การประเมินลักษณะจากการวัดสหสัมพันธ์ หรือการวัดที่คล้ายคลึงกันของลักษณะเงื่อนไขกับลักษณะตัดสินใจ การวัดนี้สามารถทำนายค่าของลักษณะหนึ่งจากลักษณะอื่นที่ใกล้เคียงได้ โดยใช้ค่า correlation coefficient (cc) หากค่า cc สูงจะสามารถจำแนกกลุ่มได้ดี

Consistency Measures การประเมินลักษณะจากการวัดความน่าเชื่อถือของกลุ่มข้อมูล และการใช้ Min-Features bias ในการเลือกลักษณะย่อย การวัดนี้พยายามหาจำนวนลักษณะที่น้อยที่สุดซึ่งเป็นตัวแบ่งแยกกลุ่มที่มีความสอดคล้องจากลักษณะทั้งหมด โดยความสอดคล้องนั้นถูกกำหนดโดย 2 instanceที่มีลักษณะเหมือนกันแต่ต่างกลุ่มกัน

Information Measures ประเมินจากความรู้ทฤษฎีสารสนเทศ ซึ่งมีการคำนวณไม่ซับซ้อน

Gini Index [6] การประเมินค่าที่บ่งบอกว่าลักษณะใดเหมาะสมเป็นลักษณะโดยวัดจากค่าความไม่บริสุทธิ์ ในแต่ละลักษณะ แล้วทำการเปรียบเทียบกับลักษณะอื่นๆเพื่อหาลักษณะที่มีค่า Gini ที่น้อยที่สุดเป็นลักษณะสำคัญ

Information Gain [7, 8] ใช้จากความรู้ทฤษฎีสารสนเทศ จะพิจารณาจากค่าความน่าจะเป็นของแต่ละลักษณะที่เป็นไปได้แล้ววัดค่าความไร้ระเบียบ (Entropy) เพื่อคัดเลือกถ้าลักษณะใดให้ค่าเกนที่สูงสุด แสดงว่าลักษณะนั้นสามารถจำแนกกลุ่มได้ดีที่สุดซึ่งInformation Gain มีการคำนวณที่ไม่ซับซ้อนแต่ยังมีความเอนเอียงในการประเมิน

Gain Ratio Criterion [9] ค่ามาตรฐานอัตราส่วนเกนเป็นวิธีประเมินลักษณะที่ใช้วิธีเช่นเดียวกับ Information Gain แต่เพิ่มการหาค่าสารสนเทศการแบ่งแยก (Split Information) [10] เพื่อแก้ไขความเอนเอียงซึ่งจะมีความละเอียดมากขึ้นโดยวิธีการคำนวณค่ามาตรฐานอัตราส่วนเกนดังนี้
วัดค่าความไร้ระเบียบของข้อมูลดังสมการ

$$E(S) = -\sum_{c=1}^{N} p(S_c) \times \log_2 p(S_c) \quad (2)$$

วัดค่า Information Gain เพื่อสร้างลำดับ ดังสมการ

$$Gain(S,V) = E(S) - \sum_{v \in value(v)} \frac{|S_v|}{S} \times E(S_v) \quad (3)$$

ลดค่าความเอนเอียง ดังสมการ

$$SplitInfo(S,V) = \sum_{i=1}^{m} - \frac{|S_i|}{|S|} \times \log_2 \frac{|S_i|}{S} \quad (4)$$

หาค่า Gain Ratio Criterion ด้วยค่าการ Split Info ดังสมการ

$$GainRatio(S,V) = \frac{Gain(S,V)}{SplitInfo(S,V)} \quad (5)$$

สำหรับงานวิจัยนี้ได้ประยุกต์ผลลัพธ์ที่ได้จากสมการที่ 5 เป็นฟังก์ชันความเหมาะสมของวิธีเจเนติกเพื่อวัดค่าความเหมาะสมเฉลี่ยของประชากรตัวนั้นๆ หากว่าค่าเฉลี่ยของประชากรตัวดังกล่าวมีค่าสูงกว่าประชากรตัวอื่นๆ แสดงว่ากลุ่มลักษณะที่คัดเลือกนั้นมีความสำคัญ ดังสมการที่ 6

$$fn = \frac{\sum GainRatio(S,V)}{V} \quad (6)$$

เมื่อ S คือลักษณะตัดสินใจ และ V คือ ลักษณะเงื่อนไข
fn คือฟังก์ชันความเหมาะสม

การคัดเลือกลักษณะโดยวิธีเจเนติก

งานวิจัยนี้นำเสนอการเลือกลักษณะโดยวิธีเจเนติกเพื่อนำไปสู่การจำแนกประเภทที่มีความแม่นยำสูงขึ้นอย่างมีนัยสำคัญโดยโดยมีวิธีการดังนี้

1. สร้างประชากรเริ่มต้นโดยวิธีการสุ่มเลือกประชากรครึ่งหนึ่งจากประชากรที่เป็นไปได้ทั้งหมด

2. คำนวณหาฟังชันก์ความเหมาะสมของประชากรเริ่มต้นแต่ละตัวจากนั้นทำการวัดการกระจายตัวของฟังชันก์ความเหมาะสมด้วยส่วนเบี่ยงเบนมาตรฐาน (standard deviation)

3. ถ้าการกระจายตัวของฟังชันก์ความเหมาะสมของประชากรแต่ละตัวมากกว่า 0.0001 จะเข้าสู่กระบวนการการไขว้เปลี่ยนและกระบวนการกลายพันธุ์เพื่อสร้างเป็นประชากรรุ่นต่อไป แต่ถ้าการกระจายตัวของฟังชันก์ความเหมาะสมน้อยกว่า 0.0001 แสดงว่ากระบวนการค้นหาคู่เข้าสู่คำตอบจะทำการหยุดหรือจะทำการหยุดเมื่อผลิตประชากรครบรุ่นที่กำหนดไว้

4. ค่าพารามิเตอร์ในการปรับเปลี่ยนสำหรับการไขว้เปลี่ยน 0.2 และ การกลายพันธุ์ 0.01 ขนาดประชากรครึ่งหนึ่งของประชากรทั้งหมด

ตัวอย่าง

สมมุติลักษณะเงื่อนไขที่เป็นไปได้ในชุดข้อมูลWeather เท่ากับ $2^n - 1$ ซึ่งมีลักษณะเงื่อนไข 4 ลักษณะV={outlook, temperature, humidity, windy} ดังนั้นโอกาสที่เป็นไปได้ของประชากรที่เกิดขึ้นดังตารางที่ 1

ตารางที่ 1 โอกาสของประชากรทั้งหมดที่เป็นไปได้

1000	0001	1001	0011	1011
0100	1100	0110	1110	0111
0010	1010	0101	1101	1111

ทำสร้างประชากรโดยวิธีสุ่มเพียง $\frac{n}{2} - 1$ ของประชากรทั้งหมด ดังนั้นโอกาสการเกิดประชากรเริ่มต้นจะได้ 6 ตัวโดยไม่คำนึงถึงโอกาสการเกิดซ้ำดังตารางที่ 2

ตารางที่ 2 การสร้างประชากร

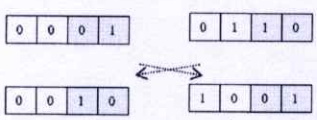
0001	1001	1100
1100	0110	1010

คำนวนความเหมาะสมด้วยฟังก์ชันหาความเหมาะสมเฉลี่ยของประชากรแต่ละตัว ดังตารางที่ 3

ตารางที่ 3 ค่าความเหมาะสมเฉลี่ยของประชากรแต่ละตัว

0001	0.15642756242117517
1100	0.269645931096278685
1001	0.180851870176935495
0110	0.207600838735547015
1100	0.269645931096278685
1010	0.172596251271031085

แล้วทำการวัดการกระจายตัวของค่าเฉลี่ยความเหมาะสมด้วยส่วนเบี่ยงเบนมาตรฐานถ้าการกระจายของค่าเฉลี่ยความเหมาะสมมากกว่า0.0001 เข้าสู่กระบวนการไขว้ เปลี่ยนต่อไปโดยการสุ่มจับคู่ ดังรูปที่ 2



รูปที่ 2 การไขว้เปลี่ยน

จะเข้าสู่กระบวนการกลายพันธุ์เมื่อความน่าจะเป็นเท่ากับ 0.01 โดยกำหนดให้กระบวนการกลายพันธุ์จะเลือกสุ่มตำแหน่งกลับบิตในสายโครโมโซมนั้น 1 ตำแหน่งตัวอย่าง ดังรูปที่ 3

รูปที่ 3 การกลายพันธุ์

ได้ประชากรรุ่นใหม่แล้ววนเข้าสู่ขั้นตอนที่ 2 – 4 จนกว่าจะได้ค่ากระจายของค่าเฉลี่ยความเหมาะสมน้อยกว่า 0.0001 หรือรุ่นประชากรที่กำหนดไว้คือ 500 รุ่น

วิธีการทดลอง

ในการทดลองนี้ได้ใช้ชุดข้อมูลจาก UCI Machine Learning Repository[11] จำนวน 6 ชุดข้อมูล ได้แก่ Heart-Statlog/(HS),/Labor,/CongressVoting/(CV), Wisconsin-Breast-Cancer(WBC) , และ Glass

ตารางที่ 4 ชุดข้อมูลการทดลอง

Dataset	No. Attribute	Attribute Type	No. Instance
HS	13	Categorical, Real	270
Labor	16	Integer,Real	57
CV	16	Categorical	435
WBC	9	Integer	699
Glass	9	Real	214

ตารางที่ 5 จำนวนลักษณะจากการคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก

Dataset	Attribute
HS	3
Labor	3
CV	7
WBC	7
Glass	6

จากตารางที่ 5 เมื่อคัดเลือกลักษณะแล้วจำนวนลักษณะที่เหลือเท่ากับ 3,3,7,7,6 ตามลำดับ จากนั้นนำชุดข้อมูลที่ได้จากการคัดเลือกไปจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Back-propagation) โดยงานวิจัยนี้ใช้การทดลองแบบ 10 - fold cross-validation แบ่งชุดข้อมูลสอนและชุดข้อมูลทดสอบ ดังนั้นการทดลองแต่ละรอบจะมีการคำนวณหาความแม่นยำ ดังสมการที่ 7

$$accuracy_i = \frac{Correctly\ Classified\ Instance}{Total\ Number\ of\ Instance} \times 100 \qquad (7)$$

และหาค่าเฉลี่ยของความแม่นยำของจำนวนรอบทั้งหมดเพื่อนำผลลัพธ์มาเปรียบเทียบประสิทธิภาพต่อไป

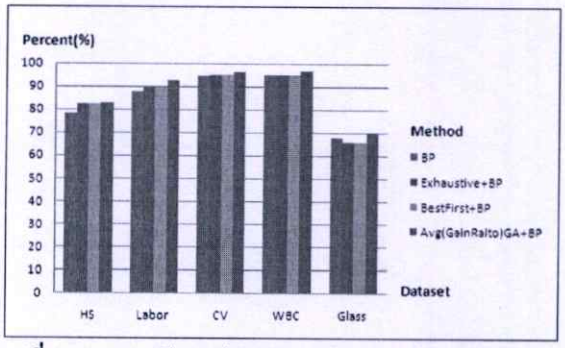
ผลการทดลอง

จากการทดลองปัญหาการจำแนกประเภทโดยการคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภท โดยใช้ขั้นตอนวิธีเจเนติกแล้วนำเข้าสู่โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับเพื่อเปรียบเทียบความแม่นยำหลังจากทำการคัดเลือก กับความแม่นยำก่อนทำการคัดเลือก ผลการเปรียบเทียบ , วิธีExhaustive และ วิธี BestFirst ดังแสดงตารางที่ 6

ตารางที่ 6 ผลการเปรียบเทียบความแม่นยำของโครงข่ายประสาทเทียมโดยการคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก

Dataset	BP	Exhaustive+ BP	BestFirst +BP	Avg(GainRaito)GA+BP
	Correctly	Correctly	Correctly	Correctly
HS	78.15 %	82.22 %	82.22 %	82.96 %
Labor	87.72 %	89.47 %	89.47 %	92.98 %
CV	94.48 %	94.94 %	94.94 %	96.32 %
WBC	95.28 %	95.28 %	95.28 %	96.85 %
Glass	67.75 %	65.89 %	65.89 %	70.09 %

จากตารางที่ 6 การจำแนกข้อมูลโดยโครงข่ายประสาทเทียมหลังการโดยการคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติกพบว่ามีความแม่นยำในการจำแนกที่สูงขึ้นเมื่อเปรียบเทียบกับก่อนการคัดเลือกลักษณะ, วิธี Exhaustive และ วิธี BestFirst



รูปที่ 4 ผลการเปรียบเทียบความแม่นยำของโครงข่ายประสาทเทียมโดยการคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติก

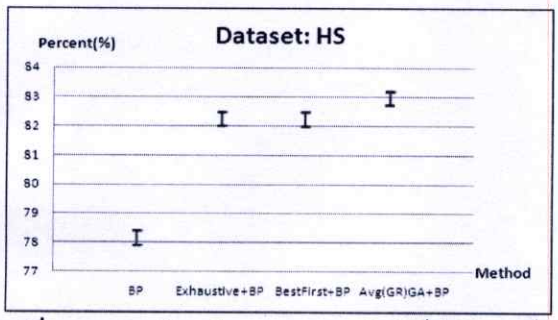
จากการทดลองโดยการคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติกพบว่าบางชุดข้อมูล เช่น ชุดข้อมูล Heart-Statlog (HS) มีลักษณะที่ให้ค่าความแม่นยำในการเรียนรู้ของโครงข่าย

เท่ากันที่ 82.96% จะพิจารณาจากค่าฟังก์ชันความเหมาะสมที่มากที่สุด เช่นเดียวกันกับกรณีชุดลักษณะที่มีจำนวนลักษณะเท่ากันและค่าความแม่นยำในการเรียนรู้ของโครงข่ายเท่ากัน แต่ลักษณะไม่ตรงกันก็จะพิจารณาที่ฟังก์ชันความเหมาะสมก่อนเสมอดังตารางที่ 7 จะทำการเลือกลักษณะที่ 1 ซึ่งมีค่าฟังก์ชันความเหมาะสมเฉลี่ยเท่ากับ 0.14177475816318574 มีจำนวน 3 แอตทริบิวต์

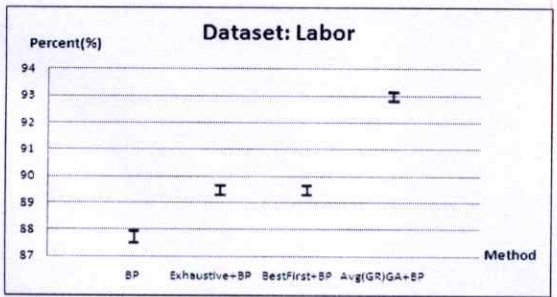
ตารางที่ 7 การคัดเลือกลักษณะด้วยค่าอัตราส่วนเกนสำหรับการจำแนกประเภทโดยใช้ขั้นตอนวิธีเจเนติกของชุดข้อมูล Heart-Statlog (HS)

No.	Attribute	Fitness Function
1	0000000010011	0.14177475816318574
2	1010100010111	0.10585499068636509
3	0010000011011	0.11902937985795307

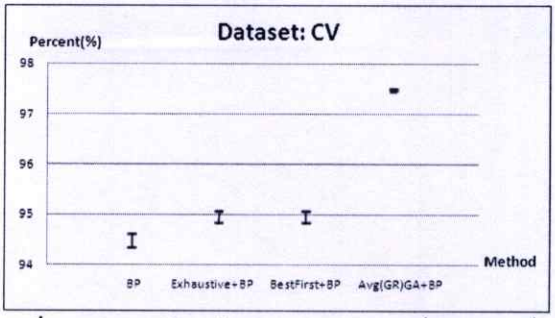
จากการทดลองแสดงผลการเปรียบเทียบของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของการเรียนรู้แบบโครงข่ายประสาทเทียม, Exhaustive+BP, BestFirst+BP และ Avg (GainRaito)GA+BP ดังรูป



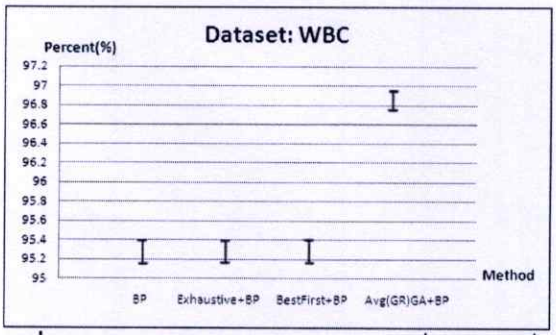
รูปที่ 5 ผลการเปรียบเทียบของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Heart-Statlog



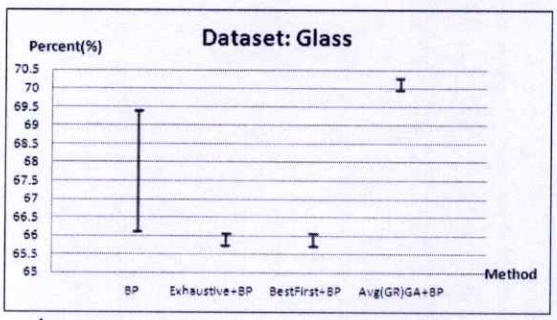
รูปที่ 6 ผลการเปรียบเทียบของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Labor



รูปที่ 7 ผลการเปรียบเทียบของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล CongressVoting



รูปที่ 8 ผลการเปรียบเทียบของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Wisconsin-Breast-Cancer



รูปที่ 9 ผลการเปรียบเทียบของความแม่นยำที่ช่วงความเชื่อมั่นระดับ 95% ของชุดข้อมูล Glass

บทสรุปและข้อเสนอแนะ

งานวิจัยนี้ทำการคัดเลือกลักษณะที่สำคัญโดยวิธีเจเนติกซึ่งร่วมกับการประเมินลักษณะโดยอัตราส่วนเกนเฉลี่ยเป็นฟังก์ชันความเหมาะสมจากผลการทดลองพบว่าการจำแนกประเภทข้อมูลด้วยลักษณะที่ได้ให้ความแม่นยำและการเรียนรู้สูงขึ้นเมื่อเปรียบเทียบกับการค้นหาลักษณะด้วยวิธี BestFirst , วิธีExhaustive และก่อนการคัดเลือกลักษณะอย่างมีนัยสำคัญของชุดข้อมูลที่นำมาทดสอบจำนวน 5 ชุดข้อมูล

ในการทดลองพบว่าค่าฟังก์ชันที่ใช้เป็นเกณฑ์ในการค้นหาและเลือกลักษณะมีผลต่อการคัดเลือกลักษณะที่ดี เนื่องจากว่าค่าฟังก์ชันเป็นตัวแปรสำคัญที่จะทำให้ค้นพบ

ลักษณะที่ดีหรือลักษณะที่ไม่เหมาะสมได้ งานวิจัยครั้งต่อไปมุ่งศึกษา วิธีการประเมินลักษณะวิธีอื่นๆ เช่น CFS (Correlation – based Feature Selection) เพื่อเปรียบเทียบกับวิธีประเมินค่ามาตรฐานอัตราส่วนเกน (Gain Ratio Criterion) และวิธีการค้นหาแบบสุ่มวิธีอื่นๆ เช่น วิธีการจาลองการอบเหนียว (Simulated Annealing) เพื่อหลีกเลี่ยงจากการติด local minima และ เปรียบเทียบการจาแนกประเภทข้อมูลด้วยวิธีโครงข่ายประสาทเทียมตัวอื่นๆต่อไป เช่น RBF และ SVM เป็นต้น

เอกสารอ้างอิง

[1] Huan Liu, Senior Member,2005, "Toward Integrating Feature Selection Algorithm for Classification and Clustering", IEEE Transactions Knowledge And Data Engineering, VOL.17, NO.4.

[2] Ron Kohavi,George H.John , 1997, "Wrappers for Feature Subset Selection", AJI special.

[3] ฐิติมา เกษมศรีธนาวัฒน์ธนัสนี เพียรตระกูล , ,2554, "การจำแนกความคิดเห็นโดยใช้ตัวจำแนกแบบเบย์ร่วมกับการคัดเลือกคุณลักษณะด้วยอัลกอริทึมรีลีฟ", CIT2011 & UniNOMS 2011.

[4] Uros Pompe,Igor Kononenko, "Linear Space Induction in First Order Logic with RELIEFF".

[5] M.Dash,H.Liu,1997, "Feature Selection for Classification" Intelligent Data Analysis,131-156.

[6] Laura E,Raileanu ,Kilian Stoffel, 2004 , "Theoretical Comparison between the Gini Index and Information Gain",Swiss National Science Foundation 41,77–93.

[7] นฤพนต์ ว่องประชานุกูล, 2548, "วิธีที่เหมาะสำหรับการตัด กิ่งต้นไม้ตัดสินใจของการทำเหมืองข้อมูล", วิทยานิพนธ์ปริญญาวิศวกรรมศาสตร์มหาบัณสาขาวิชาวิศกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

[8] Matthew N. Anyanwu, Sajjan G. Shiva, "Comparative Analysis of Serial Decision Tree Classification Algorithms ", International Journal of Computer Science and Security, Issue(3),230-240.

[9] 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.

[10] อุบลวรรณ กิจคณะ,2553, "การเรียนรู้รูปแบบรหัสพันธุกรรมเพื่อจำแนกชนิดของมะเร็งเม็ดเลือดขาวด้วยอัลกอริทึม C5.0,วารสารวิจัย มข.(บศ), 10(3),21-28

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

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวบุญญาพร เข็มปัญญา
วัน เดือน ปี	14 กันยายน 2528
ที่อยู่	22 หมู่ 4 ต.ลานตากฟ้า อ.นครชัยศรี จ.นครปฐม 73120
E-mail	KHEMPANYA.B@hotmail.com
ประวัติการศึกษา	2550 จบการศึกษาปริญญาวิทยาศาสตรบัณฑิต สาขาระบบสารสนเทศคอมพิวเตอร์ มหาวิทยาลัยบูรพา วิทยาเขตสารสนเทศจันทบุรี