# ขั้นตอนวิธีการแบ่งช่วงข้อมูลที่มีประสิทธิภาพสำหรับเหมืองข้อมูล

# Efficiency Discretization Algorithm for Data Mining

**อับดุลเลาะ บากา**

**Abdulloh Baka**

**วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา**
**วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์**
**มหาวิทยาลัยสงขลานครินทร์**

**A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of**
**Master of Science in Computer Science**
**Prince of Songkla University**
**2557**

**ชื่อวิทยานิพนธ์** ขั้นตอนวิธีการแบ่งช่วงข้อมูลที่มีประสิทธิภาพสำหรับเหมืองข้อมูล
**ผู้เขียน** นายอับดุลเลาะ บากา
**สาขาวิชา** วิทยาการคอมพิวเตอร์

---

**อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก**

.....
(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธ์)

**อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม**

.....
(ผู้ช่วยศาสตราจารย์ ดร.ศิริรัตน์ วณิชโยบล)

**คณะกรรมการสอบ**

.....ประธานกรรมการ
(ดร.นพมาศ ปักเข็ม)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธิ์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ศิริรัตน์ วณิชโยบล)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ลัดดา ปรีชาวีรกุล)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

.....
(รองศาสตราจารย์ ดร.ธีระพล ศรีชนะ)
คณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

ลงชื่อ.....

(ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธิ์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ลงชื่อ.....

(ผู้ช่วยศาสตราจารย์ ดร.ศิริรัตน์ วณิชโยบล)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม

ลงชื่อ.....

(นายอับดุลเลาะ บากา)

นักศึกษา

ขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ.....

(นายอับดุลเลาะ บากา)

นักศึกษา

**ชื่อวิทยานิพนธ์** ขั้นตอนวิธีการแบ่งช่วงข้อมูลที่มีประสิทธิภาพสำหรับเหมืองข้อมูล
**ผู้เขียน** นายอับดุลเลาะ บากา
**สาขาวิชา** วิทยาการคอมพิวเตอร์
**ปีการศึกษา** 2556

## บทคัดย่อ

ขั้นตอนวิธีในการแบ่งช่วงข้อมูลมีบทบาทสำคัญในการทำเหมืองข้อมูล เนื่องจากช่วยให้ผู้ใช้เข้าใจข้อมูลได้ง่ายขึ้น ลดเวลาในการประมวลผล ลดความซับซ้อนของข้อมูล และเพิ่มประสิทธิภาพในการทำงาน อีกทั้งยังส่งผลให้ค่าความถูกต้องสูงขึ้น วิทยานิพนธ์นี้ได้นำเสนอขั้นตอนวิธีใหม่ในการแบ่งช่วงข้อมูล โดยประยุกต์ใช้ตาราง 2D Quanta Matrix ในการหาค่าการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ในแต่ละช่วงข้อมูล ซึ่งเป็นการพิจารณาค่าความสัมพันธ์ของข้อมูลจากทั้ง 2 มิติ คือ คลาสและแอตทริบิวต์ โดยใช้ค่าเฉลี่ยการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ในทุกๆช่วงข้อมูล เรียกว่า Class Attribute Interval Average (CAIA) เป็นเกณฑ์ในการพิจารณาการหลอมรวมของช่วงข้อมูลที่อยู่ติดกัน เพื่อให้ได้ช่วงข้อมูลที่ดีที่สุด และใช้จำนวนของคลาสเป็นเกณฑ์ในการหยุด งานวิจัยนี้ได้เปรียบเทียบขั้นตอนวิธีในการแบ่งช่วงข้อมูลของ CAIA กับ 6 ขั้นตอนวิธีในการแบ่งช่วงข้อมูลคือ 1) Equal-Width 2) Equal-Frequency 3) ChiMerge 4) Information Entropy Maximization 5) Class Attribute Interdependence Maximization และ 6) Class Attribute Contingency Coefficient โดยทดสอบกับ 4 ชุดข้อมูล Benchmarks จาก UCI คือ 1) Iris 2) Breast Cancer 3) Heart Disease และ 4) Glass ใช้ตัวจำแนกประเภทข้อมูล 4 แบบ คือ J48 RBF MLP และ NB มีการทดสอบแบบ 10 Fold Cross Validation ซึ่งผลการทดลองที่ได้แสดงให้เห็นว่าอัลกอริทึมของ CAIA มีประสิทธิภาพที่ดีที่สุดทั้งจำนวนของช่วงข้อมูลที่น้อย และค่าความถูกต้องที่สูงที่สุด

**Thesis Title** Efficiency Discretization Algorithm for Data Mining

**Author** Abdulloh Baka

**Major Program** Computer Science

**Academic Year** 2013

## ABSTRACT

Discretization algorithm has an important role for data mining preprocessing. Discretization algorithm will help user to easily understand the data, reduce the complexity of data, reduce processing time, increase efficiency and accuracy of the data set. This paper proposed the new discretization algorithm called Class Attribute Interval Average (CAIA). The algorithm applied 2D quanta matrix table and calculated each individual interval's average to merge the adjacent intervals. The stopping criteria was the number of classes. The experiment used data from four UCI data sets (Iris, Breast Cancer, Heart Diseases and Glass). The experimental results from the comparison with the other six discretization algorithms (EW, EF, ChiMerge, IEM, CAIM and CACC) showed that the proposed CAIA had the best mean rank for both the number of intervals and the accuracy from all four classification algorithms (J48, RBF, MLP and NB).

# กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความช่วยเหลือและสนับสนุนจากบุคคลหลายฝ่าย ผู้วิจัยรู้สึกซาบซึ้งและขอบพระคุณอย่างสูง คือ

ผู้ช่วยศาสตราจารย์ ดร.วิภาดา เวทย์ประสิทธ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำปรึกษาแนะนำและช่วยเหลือในการแก้ปัญหาต่างๆ ให้แก่ผู้วิจัยเสมอมา พร้อมทั้งตรวจทานและแก้ไขวิทยานิพนธ์ให้แก่ผู้วิจัย

ผู้ช่วยศาสตราจารย์ ดร.ศิริรัตน์ วณิชโยบล อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วมที่กรุณาให้ข้อเสนอแนะต่างๆ รวมทั้งตรวจทานและแก้ไขวิทยานิพนธ์ให้แก่ผู้วิจัย

ดร.นพมาศ ปักเข็ม ประธานกรรมการในการสอบวิทยานิพนธ์ที่กรุณาช่วยตรวจทานและแก้ไขวิทยานิพนธ์ให้มีความสมบูรณ์

ผู้ช่วยศาสตราจารย์ ดร.ลัดดา ปรีชาวีรกุล กรรมการในการสอบวิทยานิพนธ์ที่กรุณาตรวจทานและแก้ไขวิทยานิพนธ์ให้มีความสมบูรณ์

อาจารย์ภาควิชาวิทยาการคอมพิวเตอร์ทุกท่าน ที่ให้ความรู้ด้านวิชาการซึ่งสามารถนำมาใช้ในการทำวิทยานิพนธ์ได้เป็นอย่างดี

เจ้าหน้าที่ภาควิชาวิทยาการคอมพิวเตอร์และเจ้าหน้าที่บัณฑิตวิทยาลัยทุกท่านที่ให้ความช่วยเหลือและอำนวยความสะดวกเกี่ยวกับเอกสารต่างๆ

เพื่อนๆ พี่ๆ และน้องๆ ภาควิชาวิทยาการคอมพิวเตอร์ ที่คอยให้กำลังใจ และช่วยเหลือให้คำปรึกษาในการทำวิทยานิพนธ์

คุณพ่อ คุณแม่ น้องสาว ภรรยา รวมถึงญาติๆ ทุกคน ที่ให้การสนับสนุนคอยเป็นห่วงสุขภาพและให้กำลังใจแก่ผู้วิจัยมาโดยตลอด

ผู้วิจัยขอขอบคุณทุกท่านเป็นอย่างสูงมา ณ โอกาสนี้

อับดุลเลาะ บากา

# สารบัญ

# รายการตาราง

**ตาราง** **หน้า**

**ตาราง** **หน้า**

# รายการภาพประกอบ

**ภาพประกอบ** **หน้า**

**ภาพประกอบ** **หน้า**

# บทที่ 1

# บทนำ

## 1.1 ความสำคัญและที่มาของปัญหา

เหมืองข้อมูล (Data Mining) หมายถึงกระบวนการของการกลั่นกรองสารสนเทศที่ซ่อนอยู่ในคลังข้อมูลหรือฐานข้อมูลขนาดใหญ่ เพื่อทำนายแนวโน้มและพฤติกรรมของข้อมูล โดยอาศัยข้อมูลในอดีต เพื่อค้นหากฎความสัมพันธ์ (Association Rule) จดจำรูปแบบ (Pattern Recognition) เพื่อใช้สนับสนุนการตัดสินใจ คุณลักษณะโดยทั่วไปของเหมืองข้อมูลสามารถแบ่งออกเป็น 3 คุณลักษณะใหญ่ๆ คือ 1) การเรียนรู้แบบมีผู้สอน (Supervised Learning) คอมพิวเตอร์จะทำการเรียนรู้เพื่อแบ่งกลุ่มข้อมูลจากชุดข้อมูลตัวอย่างที่ใช้ในการสอน ตัวอย่างเช่น ต้นไม้ตัดสินใจ (Decision Trees) โครงข่ายประสาทเทียม (Artificial Neural Network) และนาอีฟเบย์ (Naïve Bayes) 2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) คอมพิวเตอร์จะทำการเรียนรู้เพื่อแบ่งกลุ่มข้อมูลด้วยตัวเอง ตัวอย่างเช่นแผนที่การจัดกลุ่มเอง (Self-Organizing: SOM) การแบ่งกลุ่มเองด้วยค่า K (K-Mean) เป็นต้น และ 3) การวิเคราะห์ทางการตลาด (Market Basket Analysis) เป็นการหากฎความสัมพันธ์ (Association Rules) เช่น ถ้าลูกค้าซื้อผ้าอ้อมแล้วลูกค้ามีโอกาสที่จะซื้อเบียร์ด้วยค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) เป็นเท่าใด สามารถเขียนเป็นกฎความสัมพันธ์ได้คือ {ผ้าอ้อม}⟶ {เบียร์} เป็นต้น โดยเราสามารถนำความสัมพันธ์ที่ได้มาปรับใช้ในทางการตลาดให้สามารถขายสินค้าได้เพิ่มขึ้น หรือจัดวางชั้นสินค้าได้เหมาะสมมากขึ้น ตัวอย่างเช่น ขั้นตอนวิธีอพริโอริ (Apriori Algorithm) (Roigerand and Geatz, 2003; Ming and Xinping, 2009)

ขั้นตอนก่อนการประมวลผลข้อมูล (Data Preprocessing) คือ กระบวนการในการคัดเลือกข้อมูล เพื่อจัดรูปแบบของข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการใช้งานก่อนที่จะป้อนข้อมูลเข้าสู่กระบวนการของเหมืองข้อมูล ซึ่งเป็นขั้นตอนหนึ่งที่ใช้เวลาดำเนินการนาน ผลลัพธ์ที่ได้จากเหมืองข้อมูลจะมีคุณภาพหรือไม่ ก็ขึ้นอยู่กับขั้นตอนก่อนการประมวลผลข้อมูล ขั้นตอนก่อนการประมวลผลข้อมูลมีทั้งหมด 5 ขั้นตอน (Garcia et al., 2011; Wick, 2012) คือ 1) การรวบรวมข้อมูล (Data Integration) เป็นการรวบรวมข้อมูลจากหลายๆ แหล่งข้อมูล 2) การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนที่แก้ปัญหาค่าข้อมูลสูญหาย เช่น การแทนค่าข้อมูลสูญหายด้วย

ค่าเฉลี่ยของข้อมูลรอบข้าง หรือลบแถวข้อมูลที่มีค่าสูญหายทิ้งไป (สุคนธ์ทิพย์, 2551) 3) การสุ่มตัวอย่าง (Data Sampling) เป็นการสุ่มเลือกตัวอย่างข้อมูลบางส่วนจากข้อมูลทั้งหมดมาใช้ในการวิเคราะห์เพื่อลดปริมาณข้อมูลให้น้อยลง 4) การเลือกลักษณะเฉพาะ (Feature Selection) เป็นการเลือกเซตของลักษณะเฉพาะใหม่จากเซตของลักษณะเฉพาะเดิม โดยที่เซตของลักษณะเฉพาะใหม่ที่ได้จะเป็นเซตย่อยของเซตลักษณะเฉพาะเดิม (Liu and Setiono, 1995) และ 5) การแบ่งช่วงข้อมูล (Discretization) เป็นการแปลงค่าคุณลักษณะของข้อมูลที่มีลักษณะแบบต่อเนื่องให้อยู่ในลักษณะแบบไม่ต่อเนื่อง

โดยทั่วไปข้อมูลมี 2 ลักษณะคือ ข้อมูลแบบต่อเนื่อง (Continuous Data) เช่น ความสูง อายุ เป็นต้น และข้อมูลแบบไม่ต่อเนื่อง (Discrete Data) เช่น เพศ ระดับการศึกษา เป็นต้น (Kurgan and Cios, 2004) การเรียนรู้ของเครื่องคอมพิวเตอร์ (Machine Learning) เช่น ต้นไม้ตัดสินใจ นาอีฟเบย์ จะใช้งานกับข้อมูลที่มีลักษณะแบบไม่ต่อเนื่อง (Singh and Verma, 2009) แต่ในขณะที่ข้อมูลที่มีอยู่จริงในโลกปัจจุบันส่วนใหญ่เป็นข้อมูลที่มีลักษณะแบบต่อเนื่อง ดังนั้นจึงจำเป็นต้องมีกระบวนการในการแปลงคุณลักษณะของข้อมูลแบบต่อเนื่องให้อยู่ในรูปแบบของคุณลักษณะข้อมูลแบบไม่ต่อเนื่อง ซึ่งกระบวนการนี้เรียกว่า การแบ่งช่วงข้อมูล (Peng et al., 2009) คุณลักษณะทั่วไปของการแบ่งช่วงข้อมูลสามารถจำแนกออกเป็น 5 ประเด็น (Garcia et al., 2011) ดังนี้ 1) Supervised กับ Unsupervised 2) Multivariate กับ Univariate 3) Split กับ Merge 4) Static กับ Dynamic และ 5) Global กับ Local สำหรับข้อดีของการแบ่งช่วงข้อมูล (Peng et al., 2009 ; Sang et al., 2010) คือ 1) สามารถลดความซับซ้อนของข้อมูล ส่งผลทำให้ประหยัดพื้นที่ในการจัดเก็บข้อมูลในระบบ 2) ข้อมูลมีลักษณะเป็นแบบไม่ต่อเนื่องทำให้ผู้ใช้เข้าใจข้อมูลได้ดียิ่งขึ้น 3) ทำให้ค่าความถูกต้องในการเรียนรู้สูงขึ้น และ 4) การเรียนรู้ของคอมพิวเตอร์สามารถทำงานได้เร็วขึ้น

ปัญหาหนึ่งที่สำคัญของเหมืองข้อมูลในปัจจุบันคือ ขนาดของชุดข้อมูลที่มีขนาดใหญ่และจำนวนของแอตทริบิวต์ (Attribute) ที่มีจำนวนมาก จึงทำให้ขั้นตอนวิธีที่ใช้ในเหมืองข้อมูลมีความยุ่งยากซับซ้อนและใช้เวลานานในการสกัดความรู้จากชุดข้อมูลที่มีขนาดใหญ่ การแบ่งช่วงข้อมูลจึงเป็นอีกเทคนิคหนึ่งที่ช่วยให้ชุดข้อมูลมีขนาดเล็กลงและทำให้ผู้ใช้เข้าใจข้อมูลได้ดียิ่งขึ้น ตัวอย่างของขั้นตอนวิธีที่ใช้ในการแบ่งช่วงข้อมูล เช่น Equal-Width (EW) (Wong and Chiu, 1987), Equal-Frequency (EF) (Wong and Chiu, 1987), ChiMerge (Kerber, 1992), Information Entropy Maximization (IEM) (Fayyad and Irani, 1993), Class-Attribute Interdependent Maximization CAIM (Kurgan and Cios, 2004) และ Class-Attribute Contingency Coefficient (CACC) (Tsai, 2008) เป็นต้น จากงานวิจัยของ (Wong and Chiu, 1987; Kerber, 1992; Fayyad and

Irani, 1993; Kurgan and Cios, 2004; Tsai, 2008) แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลก่อนที่จะป้อนเข้าสู่กระบวนของเหมืองข้อมูลนั้น สามารถให้ค่าความถูกต้องที่สูงกว่า ใช้เวลาในการเรียนรู้ข้อมูลน้อยกว่า และมีประสิทธิภาพในการทำเหมืองข้อมูลที่ดีกว่า

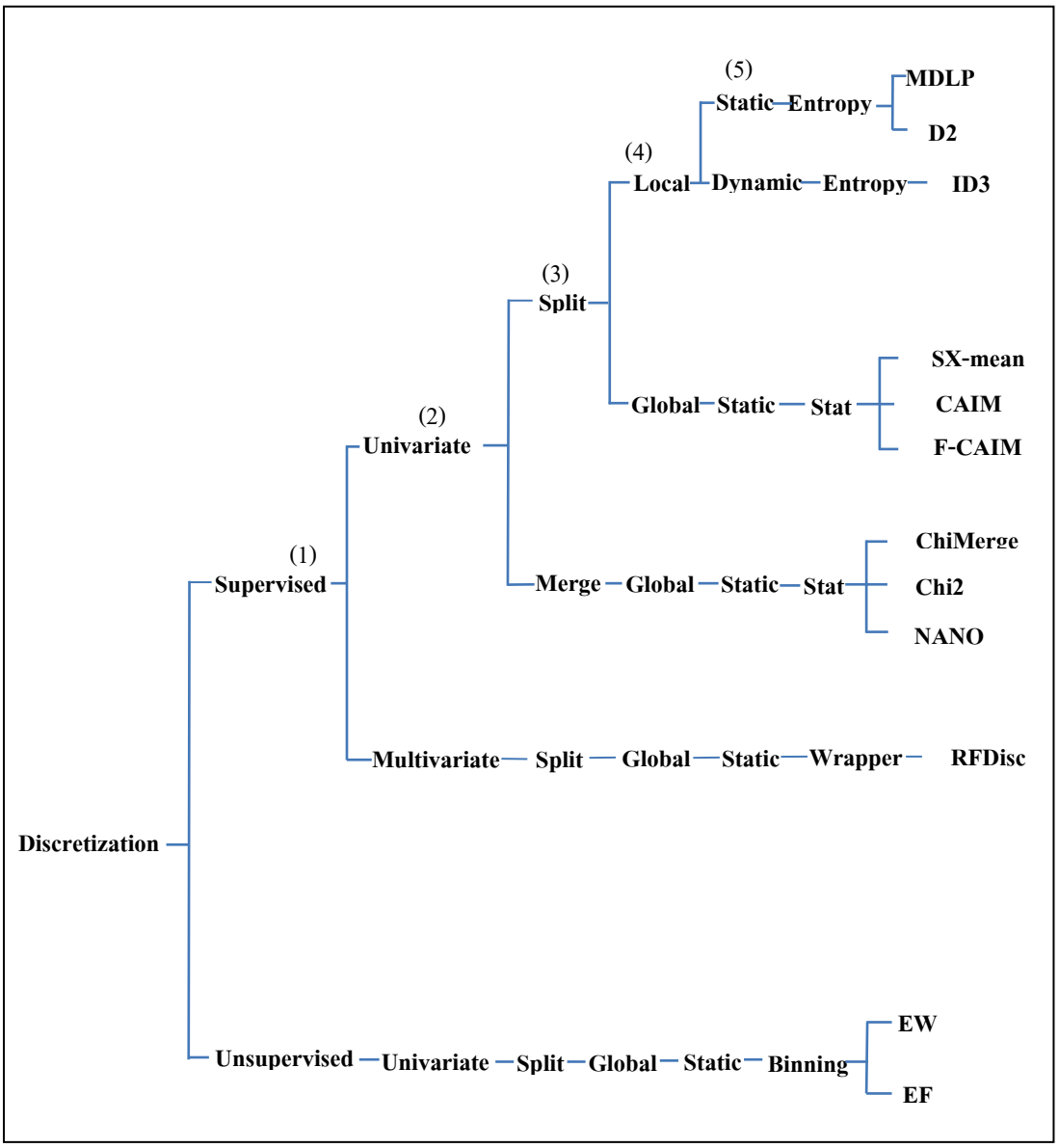
วิทยานิพนธ์ฉบับนี้ได้นำเสนอขั้นตอนวิธีในการแบ่งช่วงข้อมูลที่มีประสิทธิภาพสำหรับเหมืองข้อมูล โดยใช้ค่าเฉลี่ยการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ในทุกๆ ช่วงข้อมูล (Class Attribute Interval Average: CAIA) เป็นเกณฑ์ในการพิจารณาการแบ่งช่วงข้อมูล โดยมีคุณลักษณะเป็น supervised, univariate, merge, global และ static และใช้หลักการทางสถิติในการแบ่งช่วงข้อมูล โดยใช้ Decision Tree J48, Radial Basis Function (RBF), Multilayer Perceptron (MLP) and Naïve Bays (NB) ในการประเมินประสิทธิภาพของขั้นตอนวิธี

## 1.2 การตรวจสอบเอกสาร

เทคนิคที่ใช้ในการออกแบบขั้นตอนวิธีในการแบ่งช่วงข้อมูล (Discretization) สำหรับเหมืองข้อมูลที่มีประสิทธิภาพได้แก่ วิธีการแบ่งช่วงข้อมูล คุณลักษณะของการแบ่งช่วงข้อมูล เกณฑ์การหยุด เกณฑ์ที่ใช้ในการประเมินประสิทธิภาพ และการจำแนกประเภทข้อมูล (Peng et al., 2009)

### 1.2.1 คุณลักษณะของการแบ่งช่วงข้อมูล

คุณลักษณะโดยทั่วไปของการแบ่งช่วงข้อมูลสามารถจำแนกออกตามลักษณะลำดับชั้นแสดงดังภาพประกอบ 1.1 ได้ 5 ประเด็น (Garcia et al., 2011; Peng, 2009) คือ 1) Supervised กับ Unsupervised 2) Univariate กับ Multivariate 3) Split กับ Merge 4) Static กับ Dynamic และ 5) Local กับ Global ซึ่งแต่ละประเด็นสามารถอธิบายได้ดังต่อไปนี้




ภาพประกอบ 1.1 Hierarchical Framework of discretization Method

1) Supervised และ Unsupervised สำหรับขั้นตอนการดำเนินการของการแบ่งช่วงข้อมูลแบบมีผู้สอน (Supervised) จะมีการพิจาณากลุ่มเป้าหมาย (Target Class) ที่มีความสัมพันธ์กับคุณลักษณะ (Attribute) เพื่อใช้ในการกำหนดจุดตัดในการแบ่งช่วงข้อมูลที่ดีที่สุด ในขณะที่การแบ่งช่วงข้อมูลแบบไม่มีผู้สอน (Unsupervised) จะไม่มีการพิจารณากลุ่มเป้าหมายในการกำหนดจุดตัดของการแบ่งช่วงข้อมูล ซึ่งวิธีการแบ่งช่วงข้อมูลส่วนใหญ่ในปัจจุบันเป็นแบบมีผู้สอน

2) Univariate และ Multivariate การแบ่งช่วงข้อมูลแบบ Multivariate จะมีดำเนินการพิจารณาทุกๆ คุณลักษณะของข้อมูลแบบต่อเนื่องในเวลาเดียวกัน เพื่อกำหนดค่าเริ่มต้นในการหาจุดตัด ส่วน Univariate จะดำเนินการโดยพิจารณาคุณลักษณะของข้อมูลแบบต่อเนื่องในแต่ละรอบของการหาจุดตัดเพียงหนึ่งคุณลักษณะเท่านั้น

3) Splitting และ Merging การแบ่งแยกข้อมูล (Splitting) เป็นวิธีการในการกำหนดขอบเขต (Boundary) ของจุดตัดที่เป็นไปได้ทั้งหมดแล้วจึงทำการแบ่งแยกข้อมูลออกเป็นสองช่วง โดยจะเป็นการดำเนินงานในลักษณะจากบนลงล่าง (Top-Down) ส่วนวิธีการหลอมรวม (Merging) เริ่มต้นด้วยการกำหนดค่าของจุดตัดในการแบ่งช่วงข้อมูลก่อน แล้วจึงกำจัดจุดตัดตามเงื่อนไขที่กำหนด เพื่อหลอมรวมสองช่วงข้อมูลที่อยู่ติดกันเข้าด้วยกัน โดยจะมีการดำเนินงานในลักษณะจากล่างขึ้นบน (Bottom-Up) แต่ในบางขั้นตอนวิธีของการแบ่งช่วงข้อมูลสามารถทำงานในลักษณะที่มีทั้งสองลักษณะที่เรียกว่า Hybrid นั่นคือ มีการแบ่งแยกและหลอมรวมสลับกันไปในแต่ละช่วงของการดำเนินงาน

4) Local และ Global วิธีการแบ่งช่วงข้อมูลแบบ Global เป็นการแบ่งช่วงข้อมูลของคุณลักษณะทั้งหมดในเวลาเดียวกัน แต่วิธีการแบ่งช่วงข้อมูลแบบ Local เป็นการแบ่งช่วงข้อมูลของลักษณะข้อมูลแบบต่อเนื่องเพียงแค่หนึ่งคุณลักษณะข้อมูลเท่านั้น

5) Static และ Dynamic เป็นคุณลักษณะที่จะกล่าวถึงความเป็นอิสระของการดำเนินงานในขั้นตอนวิธีของการแบ่งช่วงข้อมูลในส่วนที่เกี่ยวข้องกับขั้นตอนวิธีของการเรียนรู้ (Learner) วิธีการแบบพลวัตร (Dynamic) จะมีการแบ่งช่วงข้อมูลในช่วงเวลาเดียวกันกับขั้นตอนของการเรียนรู้ของเครื่องคอมพิวเตอร์ แต่วิธีการแบบคงที่ (Static) จะมีการดำเนินการแบ่งช่วงข้อมูลก่อนที่จะเข้าสู่ขั้นตอนการเรียนรู้ของเครื่องคอมพิวเตอร์

### 1.2.2 เกณฑ์การหยุดที่ใช้ในการแบ่งช่วงข้อมูล

การแบ่งช่วงข้อมูลนั้นสามารถจำแนกออกเป็นประเด็นต่างๆ ตามเกณฑ์การหยุดประกอบด้วย 1) การระบุจำนวนของถัง 2) การระบุจำนวนของกลุ่มข้อมูล 3) การกำหนดค่า Chi Square Threshold 4) การหาค่าข้อมูลที่ไม่สอดคล้องกัน (Inconsistency) 5) การหาความ

ซ้ำซ้อนกันระหว่างคลาสและคุณลักษณะของข้อมูล (Class Attribute Independence Redundancy: CAIR) และ 6) การหาค่าสูงสุดที่ขึ้นตรงต่อกันระหว่างคลาสและคุณลักษณะของข้อมูล (Class Attribute Independence Maximization: CAIM) นอกจากนี้ยังสามารถจำแนกตามคุณลักษณะการทำงานของแต่ละขั้นตอนวิธีในการแบ่งช่วงข้อมูลดังแสดงในตารางที่ 1.1 ซึ่งเป็นตารางสรุปรายละเอียดของแต่ละขั้นตอนวิธีในการแบ่งช่วงข้อมูล มีรายละเอียดดังนี้

1) การระบุจำนวนของถัง เป็นเกณฑ์การหยุดที่จะต้องมีการระบุจำนวนของถัง เพื่อใช้ในการกำหนดจำนวนของช่วงข้อมูล ตัวอย่างเช่น การแบ่งช่วงข้อมูลตามความกว้างที่เท่ากันหรือแบ่งช่วงข้อมูลตามความถี่ที่เท่ากัน (Wong and Chiu, 1987) ซึ่งเป็นวิธีการเรียนรู้แบบไม่มีผู้สอน ซึ่งสองวิธีนี้ผู้ใช้จำเป็นต้องมีการระบุจำนวนของความกว้างหรือจำนวนของความถี่ในแต่ละถังที่เท่ากันก่อนที่จะมีการแบ่งช่วงข้อมูล

2) การระบุจำนวนของกลุ่มข้อมูล เป็นเกณฑ์การหยุดที่ผู้ใช้ต้องมีการะบุค่า K ก่อนที่จะมีการแบ่งช่วงข้อมูล โดยที่ค่า K คือจำนวนของช่วงข้อมูลที่ต้องการแบ่งเช่น K-mean Clustering มีคุณสมบัติเป็นการเรียนรู้แบบไม่มีผู้สอน (Hartigan and M. Wong, 1979)

3) การกำหนดค่า Chi Square Threshold เช่น Kerber (1992) ได้นำเสนอวิธีการ ChiMerge ซึ่งมีคุณสมบัติเป็นการเรียนรู้แบบมีผู้สอนและเป็นวิธีการที่ใช้ค่าทางสถิติ $x^2$ มาช่วยในการหาจำนวนช่วงข้อมูลที่ดีที่สุดในการแบ่งช่วงข้อมูล และใช้ค่า $\chi^2$-Threshold เป็นเกณฑ์ในการพิจารณาว่าจะมีการหลอมรวมต่อของช่วงข้อมูลที่อยู่ติดกันหรือไม่ วิธีนี้ผู้ใช้ต้องมีการกำหนดระดับค่านัยสำคัญ (Significance Level) ก่อนที่จะมีการแบ่งช่วงข้อมูล

4) การหาค่าข้อมูลที่ไม่สอดคล้องเนื่องจากปัญหาข้างต้นที่ผู้ใช้จำเป็นต้องมีการกำหนดค่าบางค่าก่อนที่จะมีการแบ่งช่วงข้อมูล ดังนั้น Liu และ Setiono (1995) จึงได้นำเสนอวิธีการ Chi2 เพื่อแก้ปัญหาวิธีการของ ChiMerge ที่ผู้ใช้จำเป็นต้องมีการกำหนดระดับค่านัยสำคัญด้วยตัวเอง แต่ใน Chi2 จะใช้ค่าข้อมูลที่ไม่สอดคล้องกันเป็นเกณฑ์ในการหยุด แต่เกณฑ์การหยุดจะพิจารณาแค่ช่วงของข้อมูลที่มีผลต่อค่าดีกรีอิสระ (Degree of Freedom) เท่านั้น ต่อมา Sang และคณะ (2010) จึงได้นำเสนอวิธีการ EBDA (Effective Bottom-up Discretization Algorithm) ที่พัฒนาต่อจากวิธีการของ Chi2 ซึ่งเกณฑ์การหยุดที่ใช้จะมีการพิจารณาตัวแปรของช่วงข้อมูลที่อยู่ติดกัน

ตารางที่ 1.1 สรุปวิธีการแบ่งช่วงข้อมูล

| Authors | Year | Algorithms | Stopping Criterion | Discretization method | Characteristics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Unsupervised | Supervised | Multivariate | Univariate | Global | Local | Dynamic | Static | Split | Merge |
| Wong and Chiu | 1987 | Equal-width | Fixed bin No. | binning | ✓ | × | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| Wong and Chiu | 1987 | Equal-frequency | Fixed bin No. | binning | ✓ | × | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| Kerber | 1992 | ChiMerge | Threshold | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| Fayyad and Irani | 1993 | IEM | MDLP | information entropy | × | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ | × |
| Liu and Setiono | 1995 | Chi2 | Inconsistency | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| Monti and Cooper | 1998 | Multi-Bayesian | Bayesian | information entropy | × | ✓ | ✓ | × | × | ✓ | ✓ | × | ✓ | × |
| Kurgan and Cios | 2004 | CAIM | CAIM | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| Kurgan and Cios | 2003 | F-CAIM | CAIM | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| Kang et al. | 2006 | FastICA | ICA | statistical | × | ✓ | ✓ | × | × | ✓ | × | ✓ | × | ✓ |
| Tsai et al. | 2008 | CACC | CACC | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| Hua and Zhao | 2009 | SX-Mean | BIC | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| Sang et al. | 2010 | EBDA | Inconsistency | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| Chaoqun et al. | 2011 | NCL-CAIR | CAIR | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | ✓ |
| Proposed Method | 2013 | CAIA | CAIA | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |

ทุกคู่ และมีการพิจารณาการกระจายตัวของข้อมูลด้วย ทำให้ช่วงข้อมูลที่อยู่ติดกันทุกๆ คู่มีโอกาสที่จะหลอมรวมกันที่เท่าเทียมกัน

5) การหาความซ้ำซ้อนกันระหว่างคลาสและคุณลักษณะของข้อมูลเช่น วิธีการของ NCL-CAIR (Number of Cluster – Class Attribute Interdependence Redundancy) เป็นวิธีการที่มีการพิจารณาการกระจายตัวของข้อมูลเพื่อหาจุดเริ่มต้นของจุดตัดในการแบ่งช่วงข้อมูล แล้วหาค่าความสัมพันธ์ระหว่างคุณลักษณะของข้อมูลกับคลาสโดยใช้ค่า CAIR เป็นเกณฑ์ในการหาจุดตัดและเป็นเกณฑ์ในการหยุดของการแบ่งช่วงข้อมูล (Chaoqun et al., 2011) ส่วน Ching และคณะ (1995) ได้นำเสนอวิธีการ CADD (Class-Attribute Interdependency) ซึ่งเป็นวิธีการที่ใช้การหาค่าความสัมพันธ์ที่มากที่สุดระหว่างคลาสและคุณลักษณะของข้อมูลในการแบ่งช่วงข้อมูล โดยใช้ค่า CAIR เป็นเกณฑ์ในการกำหนดจำนวนของช่วงข้อมูลที่ดีที่สุดเช่นกัน

6) การหาค่าสูงสุดที่ขึ้นตรงต่อกันระหว่างคลาสและคุณลักษณะของข้อมูล เช่น Kurgan และ Cios (2004) ได้นำเสนอวิธีการ CAIM (Class Attribute Independence Maximization) ซึ่งมีคุณสมบัติเป็น Top-Down โดยจะใช้ค่าความสัมพันธ์ที่ขึ้นตรงต่อกันของคลาสและคุณลักษณะของข้อมูลที่มากที่สุดเป็นเกณฑ์ในการแบ่งแยกข้อมูล แต่ข้อเสียของวิธีการนี้คือจำนวนของช่วงข้อมูลที่ได้จาการแบ่งช่วงข้อมูลจะใกล้เคียงกับจำนวนของคลาส และจะพิจารณาแค่คลาสที่เป็นกลุ่มใหญ่ๆ เท่านั้น ต่อมา Kurgan และ Cios (2004)ได้พัฒนาความเร็วของ CAIM และตั้งชื่อเป็น F-CAIM (Fast Class Attribute Independence Maximization) เนื่องจากจุดอ่อนของ CAIM อยู่ที่การพิจารณาในการเลือกจุดตัด โดยจุดตัดจะเริ่มต้นที่ค่ามากสุด ต่ำสุด และทุกๆ ค่ากลางของค่าที่อยู่ติดกัน ทำให้จำนวนของจุดตัดเท่ากับ M+1 ที่ M เป็นจำนวนของข้อมูลทั้งหมด แต่ F-CAIM จะเริ่มต้นในการหาจุดตัดที่ค่ามากสุด ต่ำสุด และทุกๆ ค่ากลางของคลาสที่แตกต่างกันเท่านั้น จะทำให้จุดตัดในการเริ่มต้นน้อยลง ส่งผลทำให้เวลาที่ใช้ในการทำงานเร็วขึ้นแต่ยังคงมีคุณสมบัติทุกอย่างเหมือนกับ CAIM ทุกประการ สำหรับข้อจำกัดของ CAIM มีสองข้อคือ จำนวนของช่วงข้อมูลที่ได้จะใกล้เคียงกับจำนวนของคลาส และแต่ละครั้งที่มีการแบ่งช่วงข้อมูล CAIM จะพิจารณาแค่คลาสที่มีตัวอย่างข้อมูลมากที่สุดและไม่สนใจคลาสอื่นๆ ดังนั้นทำให้ประสิทธิภาพของการแบ่งช่วงข้อมูลด้วย CAIM criterion อาจจะไม่ดีในบางกรณี ดังนั้น Tsai และคณะ (2008) ได้นำเสนอ CACC (Class-Attribute Contingency Coefficient) เพื่อแก้ปัญหาข้อจำกัดของ CAIM เนื่องจากวิธีการแบ่งช่วงข้อมูลของ CACC จะมีการพิจารณาทุกๆ ตัวอย่างข้อมูลที่มีการแบ่งช่วง

ข้อมูล โดยการหาความสัมพันธ์ระหว่างตัวแปรสองตัวแปรคือ คลาสและแอตทริบิวต์ ทำให้การเรียนรู้ของเครื่องคอมพิวเตอร์มีประสิทธิภาพในการทำงานมากขึ้น

**1.2.3 การจำแนกประเภทข้อมูล**

การจำแนกประเภทข้อมูล (Data Classification) คือการจำแนกประเภทข้อมูลให้อยู่ในประเภทข้อมูลที่กำหนดไว้ให้ได้ โดยกระบวนการในการจำแนกข้อมูลนั้นจะต้องมีการป้อนข้อมูลชุดสอน (Training Data) เพื่อให้ระบบมีการเรียนรู้ก่อน หลังจากนั้นจึงนำข้อมูลชุดทดสอบ (Test Data) ซึ่งเป็นข้อมูลอีกกลุ่มหนึ่งมาใส่ในผลลัพธ์ที่ได้สร้างแบบจำลองมาแล้ว ตัวอย่างของการจำแนกประเภทข้อมูล คือ ต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม และนาอีฟเบย์ (สุคนธ์ทิพย์, 2551) สามารถอธิบายได้ดังต่อไปนี้

**1) ต้นไม้ตัดสินใจ**

ต้นไม้ตัดสินใจ (Decision Tree) เป็นการจำแนกข้อมูลโดยแทนความรู้ในรูปแบบของต้นไม้ โดยที่แต่ละโหนด (Node) แสดงคุณลักษณะที่ใช้ในการทดสอบข้อมูล แต่ละกิ่ง (Branch) แสดงผลในการทดสอบและโหนดใบ (Leaf Node) แสดงกลุ่มหรือคลาสที่กำหนดไว้ (กาญจนา, 2551)

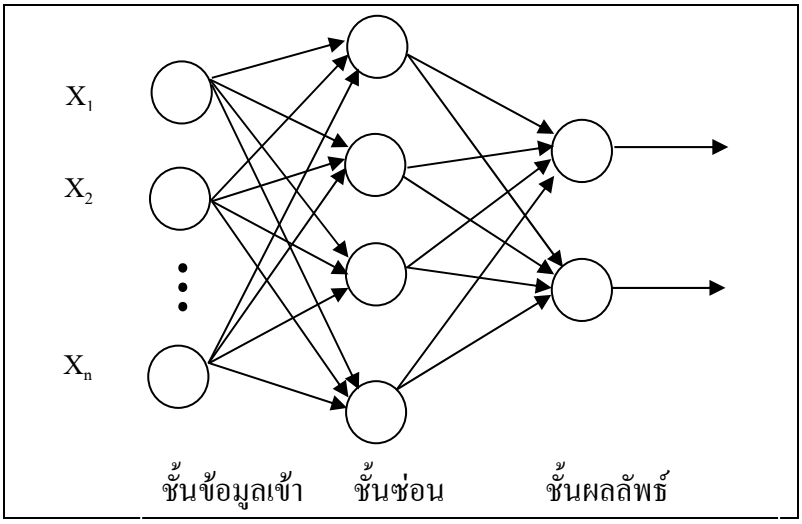
**2) โครงข่ายประสาทเทียม**

โครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นเทคนิคหนึ่งของปัญญาประดิษฐ์ ซึ่งมีความทนทานต่อความผิดพลาด (Fault Tolerant) และสามารถรองรับข้อมูลที่ไม่สมบูรณ์หรือมีสิ่งรบกวนได้ (Mellit and Kalogirou, 2008; Shiva and khare, 2004) มีรูปแบบการประมวลผลที่เลียนแบบการทำงานของเซลล์ประสาทมนุษย์ประกอบด้วยหน่วยประมวลผลย่อยเพอร์เซฟตรอน (Perceptron) หลายหน่วยเชื่อมต่อกัน องค์ประกอบของเพอร์เซฟตรอนประกอบด้วย ฟังก์ชันผลรวม (Summation Function) ทำหน้าที่หาผลรวมของผลคูณระหว่างค่าน้ำหนักข้อมูลเข้ากับค่าข้อมูลเข้า และมีฟังก์ชันกระตุ้น (Activation Function) ทำหน้าที่แปลงผลลัพธ์จากฟังก์ชันผลรวมให้อยู่ในช่วงค่าที่ต้องการ

สถาปัตยกรรมของโครงข่ายประสาทเทียมประกอบด้วย โครงข่ายประสาทเทียมแบบป้อนไปข้าง และโครงข่ายประสาทเทียมแบบเวียนกลับมีละเอียดดังต่อไปนี้

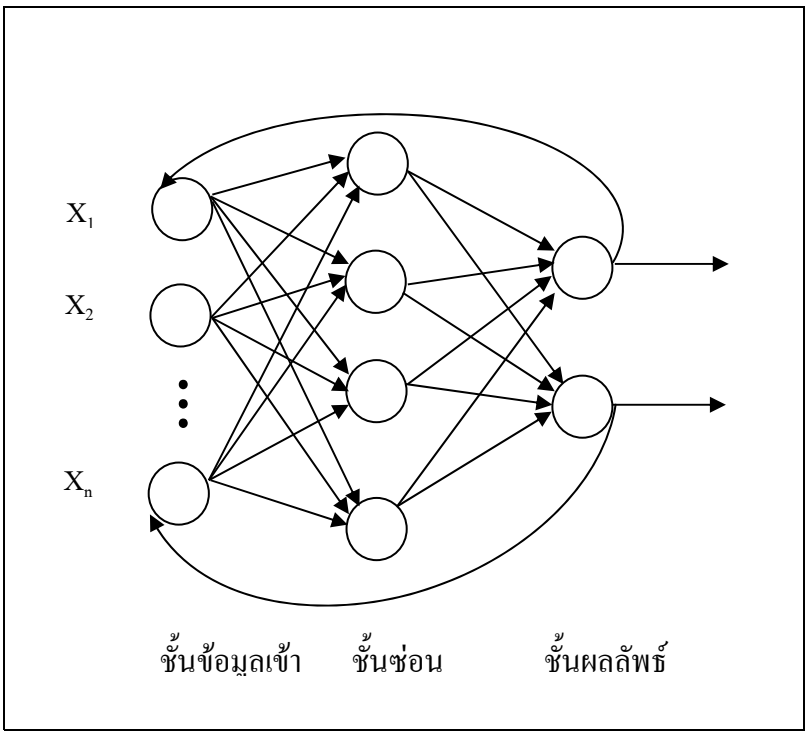
1) โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า (Feed Forward Networks) เป็นการนำเอาหน่วยประมวลผลย่อยหลายหน่วยมาเชื่อมต่อกันเป็นโครงข่ายเพื่อเพิ่ม

ประสิทธิภาพในการพยากรณ์ โดยข้อมูลที่ประมวลผลในโครงข่ายจะถูกส่งไปในทิศทางเดียวกัน จากจุดต่อด้านข้อมูลเข้า (Input Nodes) ส่งต่อไปเรื่อยๆ จนถึงจุดต่อด้านข้อมูลออก (Output Nodes) โดยไม่มีการย้อนกลับ มีการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งจะต้องมีการสอนโครงข่ายประสาทเทียมด้วยข้อมูลชุดสอน ก่อนนำไปทดสอบด้วยข้อมูลชุดทดสอบ โครงสร้างของโครงข่ายประสาทเทียมแบบหน่วยประมวลผลย่อยหลายชั้นมี 3 ระดับ คือ ชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) โดยในแต่ละชั้นนั้นจะมีหน่วยประมวลผลย่อย ซึ่งโครงข่ายประสาทเทียมแบบเพอรเซฟตรอนหลายชั้นมีประสิทธิภาพในการหาค่าความถูกต้องสูง (Wettayaprasit and Nanakorn, 2006) แสดงดังภาพประกอบ 1.2




ภาพประกอบ 1.2 โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า (Feed Forward Networks)

2) โครงข่ายประสาทเทียมแบบเวียนกลับ (Recurrent Network) เป็นโครงข่ายประสาทเทียมที่ผลลัพธ์ของเพอร์เซฟตรอนหนึ่งสามารถย้อนกลับไปเป็นข้อมูลเข้าของเพอร์เซฟตรอนในระดับก่อนหน้าได้หลายๆ ครั้ง จนกระทั่งได้คำตอบที่ต้องการ หรือที่เรียกว่าโครงข่ายแบบเวียนกลับ (Recurrent Networks) (Dorado et al., 2002) แสดงดังภาพประกอบ 1.3




ภาพประกอบ 1.3 โครงข่ายประสาทเทียมแบบเวียนกลับ (Recurrent Networks)

**3) นาอีฟเบย์**

นาอีฟเบย์ (Naïve Bayes) เป็นเทคนิคการจำแนกข้อมูลพัฒนาโดย Thomas Bayes โดยมีการตั้งสมมติฐานเพื่อกำหนดให้การเกิดของเหตุการณ์ต่างๆ ที่ใช้ในการจัดกลุ่มนั้นเป็นอิสระต่อกัน ซึ่งจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นของแต่ละความสัมพันธ์ จุดประสงค์เพื่อต้องการสร้างแบบจำลองที่อยู่ในรูปของความน่าจะเป็น เพื่อหาว่าสมมติฐานใดถูกต้องมีที่สุด ข้อดีของวิธีการแบบเบย์คือ เราสามารถใช้ข้อมูลและความรู้ที่ได้จากการเรียนรู้ก่อนหน้านี้เข้ามาช่วยในการเรียนรู้ของข้อมูลอีกด้วย นอกจากนี้ยังเหมาะสมกับชุดข้อมูลที่มีขนาดใหญ่และคุณลักษณะข้อมูลที่เป็นอิสระต่อกัน (John, Pat Langley, 1995)

## 1.3 วัตถุประสงค์ของโครงการ

1.3.1 เพื่อวิเคราะห์ ออกแบบ และพัฒนาขั้นตอนวิธีในการแบ่งช่วงข้อมูลสำหรับเหมืองข้อมูลที่มีประสิทธิภาพ

## 1.4 ขอบเขตการดำเนินงาน

1.4.1 วิเคราะห์ ออกแบบ และพัฒนาขั้นตอนในวิธีการแบ่งช่วงข้อมูลสำหรับการทำเหมืองข้อมูล

1.4.2 ชุดข้อมูลที่ใช้ในการทดลองเป็นชุดข้อมูลจาก UCI Data Set

## 1.5 ขั้นตอนการดำเนินการวิจัยและระยะเวลาการดำเนินการวิจัย

1.5.1 ศึกษางานวิจัยและเอกสารที่เกี่ยวข้องสำหรับการแบ่งช่วงข้อมูล และจำแนกข้อดีข้อเสียของแต่ละวิธีการในการแบ่งช่วงข้อมูล

1.5.2 ศึกษาเทคโนโลยีและเครื่องมือสนับสนุนสำหรับงานวิจัย

1.5.3 วิเคราะห์และออกแบบขั้นตอนวิธีที่ใช้ในการแบ่งช่วงข้อมูล

1.5.4 เขียนโปรแกรมคอมพิวเตอร์สำหรับการแบ่งช่วงข้อมูลตามที่ได้วิเคราะห์และออกแบบไว้

1.5.5 ทดสอบประสิทธิภาพของขั้นตอนวิธีที่ได้ออกแบบไว้กับขั้นตอนวิธีอื่นที่มีคุณสมบัติคล้ายๆ กัน และทดสอบกับชุดข้อมูลจาก UCL Data Set

1.5.6 เขียนบทความวิจัย

1.5.7 จัดทำเอกสารวิทยานิพนธ์

ระยะเวลาการดำเนินการวิจัยสามารถแสดงได้ดังตารางที่ 1.2

ตารางที่ 1.2 ขั้นตอนการดำเนินการวิจัยและระยะเวลาการดำเนินการวิจัย

| กิจกรรม/ขั้นตอนการดำเนินงาน | เดือน | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | พ.ศ. 2555 | | | | | | | | | พ.ศ. 2556 | | | | | | | | | | | | | พ.ศ. 2557 | | | |
| | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 1 | 2 | 3 | 4 |
| 1.ศึกษางานวิจัยและเอกสารที่เกี่ยวข้อง | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2.ศึกษาเทคโนโลยีและเครื่องมือสำหรับงานวิจัย | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3.วิเคราะห์และออกแบบขั้นตอนวิธี | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4.เขียนโปรแกรมขั้นตอนวิธี | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5.ทดสอบประสิทธิภาพ | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6.เขียนบทความวิจัย | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7.จัดทำเอกสารวิทยานิพนธ์ | | | | | | | | | | | | | | | | | | | | | | | | | |

## 1.6 สถานที่และเครื่องมือที่ใช้ในงานวิจัย

1.6.1 สถานที่ทำวิจัย

ห้องปฏิบัติการวิจัยปัญญาประดิษฐ์ (Artificial Intelligent Research LAB) ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่

1.6.2 เครื่องมือและอุปกรณ์ที่ใช้

1) ด้านฮาร์ดแวร์

เครื่องคอมพิวเตอร์ส่วนบุคคล หน่วยความจำขนาด 4GB และฮาร์ดดิสก์ความจุ 1 TB สำหรับใช้ในการพัฒนาและทดสอบระบบ

2) ด้านซอฟต์แวร์

-ระบบปฏิบัติการ Microsoft Windows 7

-โปรแกรม WEKA เวอร์ชั่น 3.7.10

-โปรแกรมภาษา C

-โปรแกรม MS Office

## 1.7 ประโยชน์ที่คาดว่าจะได้รับ

ได้ขั้นตอนวิธีที่ใช้ในการแบ่งช่วงข้อมูล (Discretization) สำหรับเหมืองข้อมูลที่มีประสิทธิภาพ

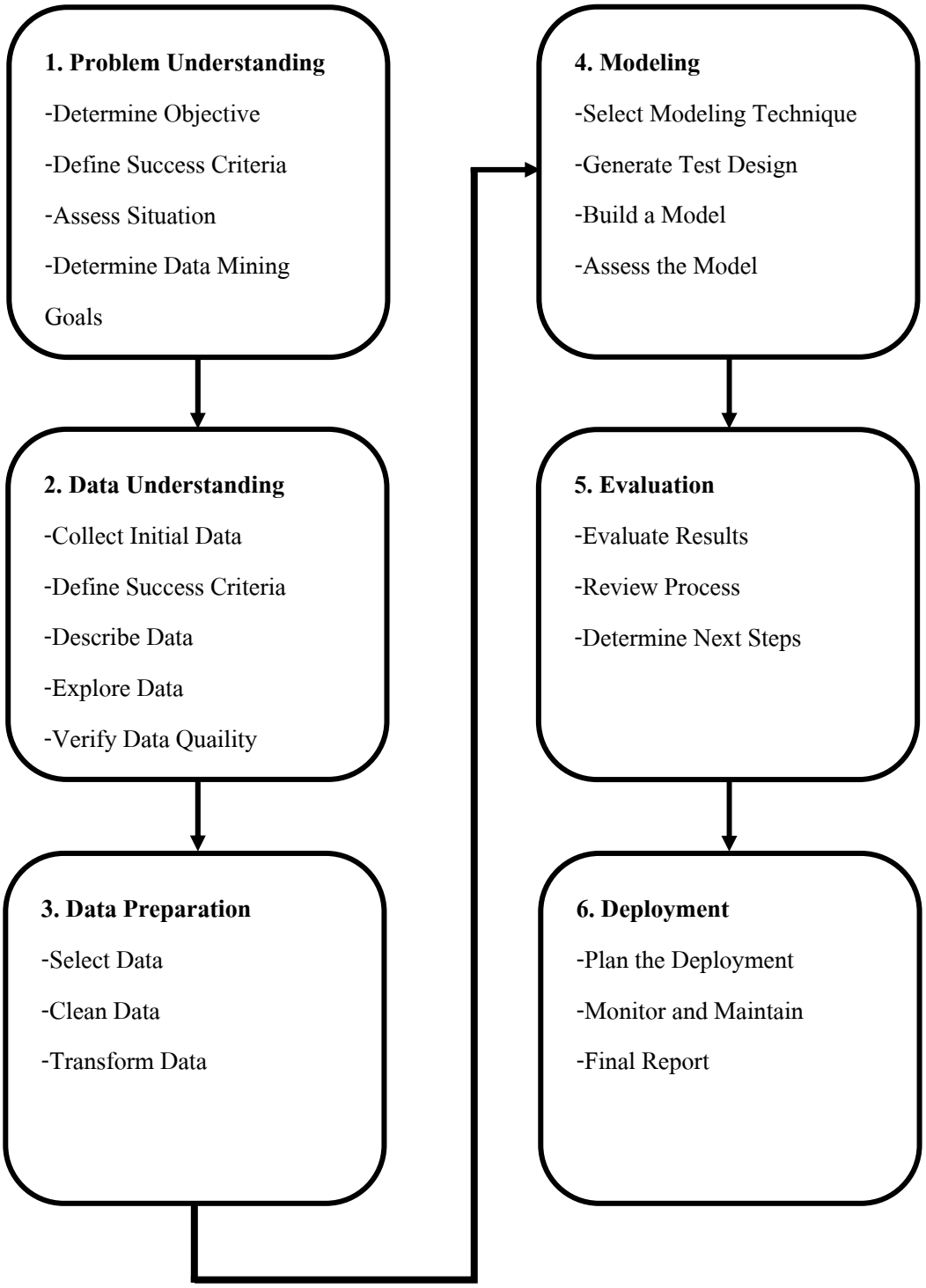
# บทที่ 2

# ทฤษฎีที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยต่างๆ ที่เกี่ยวข้องกับการพัฒนาขั้นตอนวิธีในการแบ่งช่วงข้อมูลที่ใช้สำหรับเหมืองข้อมูลที่มีประสิทธิภาพ ซึ่งประกอบด้วย 1) ทฤษฎีของการทำเหมืองข้อมูล เพื่อให้ผู้วิจัยเข้าใจความหมาย หลักการในการทำงาน ประเภทข้อมูลที่สามารถทำเหมืองข้อมูล และขั้นตอนในการทำเหมืองข้อมูล 2) ทฤษฎีการเตรียมข้อมูล เพื่อใช้ในการเตรียมข้อมูลก่อนการแบ่งช่วงข้อมูล 3) ทฤษฎีของการแบ่งช่วงข้อมูล เพื่อให้ผู้วิจัยสามารถจำแนกวิธีการ คุณลักษณะ เกณฑ์ที่ใช้ในการหยุด และเกณฑ์ที่ใช้ในการประเมินประสิทธิในการทำงานของการแบ่งช่วงข้อมูล และ 4) ทฤษฎีที่ใช้ในการจำแนกประเภทข้อมูล เช่น ต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม และนาอีฟเบย์ เพื่อใช้ในการหาค่าความถูกต้องและเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีในการแบ่งช่วงข้อมูลที่นำเสนอเพื่อใช้ในการเปรียบเทียบกับกับขั้นตอนวิธีอื่น โดยมีรายละเอียดดังต่อไปนี้

## 2.1 เหมืองข้อมูล

การทำเหมืองข้อมูลคือ กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์ รวมทั้งในด้านเศรษฐกิจและสังคม การทำเหมืองข้อมูลเป็นกระบวนการหนึ่งในการจัดเก็บและตีความหมายข้อมูลมาเป็นสารสนเทศและค้นพบความรู้ที่ซ่อนอยู่ในข้อมูลได้ นอกจากนี้ Wenke Lee และคณะ (2001) ให้ความหมายของการทำเหมืองข้อมูลว่าเป็นกระบวนการเพื่อกลั่นกรองข้อมูลจากฐานข้อมูลขนาดใหญ่ที่มีอยู่ โดยมองที่ความสัมพันธ์ของข้อมูล แนวโน้มของข้อมูลต่าง ๆ เพื่อให้สามารถนำข้อมูลที่กลั่นกรองได้นำไปใช้ประโยชน์เป็นข้อมูลสนับสนุนในการตัดสินใจ ขั้นตอนในการทำเหมืองข้อมูลประกอบด้วยหลายขั้นตอนย่อยดังนี้คือ 1) การทำความเข้าใจปัญหา 2) การทำความเข้าใจข้อมูล 3) การเตรียมข้อมูล 4) การสร้างแบบจำลอง 5) การประเมินผล และ 6) การนำไปใช้ (Roigerand and Geatz, 2003 ) ซึ่งรายละเอียดของขั้นตอนต่างๆ แสดงดังภาพประกอบ 2.1 ดังต่อไปนี้




ภาพประกอบ 2.1 ขั้นตอนการทำเหมืองข้อมูล (Roigerand and Geatz, 2003)

**2.1.1 การทำความเข้าใจปัญหา (Problem Understanding)** ประกอบด้วยกระบวนการย่อยดังนี้

- Determine Objective: การตั้งเป้าหมายในการทำเหมืองข้อมูลครั้งนี้ต้องการแก้ปัญหาใด

- Define Success Criteria: ตั้งเกณฑ์วัดความสำเร็จ อาจเป็นความสำเร็จด้านรูปธรรม เช่น เพิ่มยอดขายได้ 5% หรือในด้านนามธรรม เช่น การค้นพบความรู้ใหม่จากข้อมูล เป็นต้น

- Assess Situation: การประเมินสถานการณ์ในด้านต่าง ๆ เช่น ความพร้อมในด้านความรู้ในการจัดทำ Data Mining ความคุ้มค่าในการดำเนินการ เป็นต้น

- Determine Data Mining Goals: ตั้งเป้าหมายในเชิงการทำเหมืองข้อมูล เช่น การหาลักษณะของลูกค้าที่มีแนวโน้มซื้อสินค้า เป็นต้น

- Produce a Project Plan: วางแผนการทำเหมืองข้อมูล เช่น จะเก็บข้อมูลอย่างไร ใช้ขั้นตอนอะไร เป็นต้น

**2.1.2 การทำความเข้าใจข้อมูล (Data Understanding)** ประกอบด้วยกระบวนการย่อยดังนี้

- Collect Initial Data: การรวบรวมข้อมูลเริ่มต้น

- Define Success Criteria: กำหนดคุณสมบัติที่ต้องการเก็บ

- Describe Data: อธิบายข้อมูล

- Explore Data: สำรวจข้อมูล

- Verify Data Quality: ปรับปรุงให้ข้อมูลมีคุณภาพ

**2.1.3 การเตรียมข้อมูล (Data Preparation)** ประกอบด้วยกระบวนการย่อยดังนี้

- Select Data: การคัดเลือกข้อมูลที่จะนำมาใช้งาน โดยมีจุดประสงค์คือ การระบุแหล่งที่มาของข้อมูล และการดึงเอาข้อมูลออกมาใช้สาหรับการวิเคราะห์เบื้องต้นในการเตรียมตัวสำหรับการทำเหมืองข้อมูล การคัดเลือกข้อมูลนั้นจะแตกต่างไปตามวัตถุประสงค์ของแต่ละธุรกิจ ที่ได้กำหนดไว้ตั้งแต่ต้น และการคัดเลือกข้อมูลก็ยังถูกกำหนดโดยลักษณะงานที่จะถูกนำมาใช้อีกด้วย ตัวแปรที่ถูกเลือกมาแต่ละตัวนั้นจะต้องทำความเข้าใจว่าตัวแปร

แต่ละตัวหมายความว่าอะไร ประกอบด้วยอะไร ไม่เพียงแต่คำจำกัดความทางธุรกิจเท่านั้น แต่จะต้องมีคำอธิบายอย่างชัดเจนเกี่ยวกับชนิดของข้อมูล, ค่าที่เป็นไปได้, แหล่งกำเนิดของข้อมูล, รูปแบบของข้อมูล เป็นต้น

- Clean Data: ปรับแต่งให้ข้อมูลมีความเหมาะสมต่อการใช้งาน
- Transform Data: แปลงข้อมูลให้มีความเหมาะสมในการตัดสินใจ

**2.1.4 การสร้างแบบจำลอง (Modeling)** ประกอบด้วยกระบวนการย่อยดังนี้

- Select Modeling Technique: เลือกขั้นตอนวิธีที่เหมาะสมในการทำเหมืองข้อมูล
- Generate Test Design: กำหนดรูปแบบการทดสอบผลลัพธ์
- Build a Model: ลงมือสร้าง Model ตามขั้นตอนวิธีที่กำหนด
- Assess the Model: ทดสอบ Model ที่ได้ว่ามีความถูกต้องและน่าเชื่อถือมากน้อยเพียงใด

**2.1.5 การประเมินผล (Evaluation)** ประกอบด้วยกระบวนการย่อยดังนี้

- Evaluate Results: ประเมิน Model ที่สร้างขึ้นด้วยการทดลองนำไปใช้กับสถานการณ์จริง หรือจำลองขึ้นเพื่อดูว่า Model ที่สร้างได้ผลหรือไม่เพียงใด
- Review Process: การทบทวนตัวผลการทดสอบ Model
- Determine Next Steps: พิจารณาขั้นตอนการดำเนินงานและกระบวนการที่ผ่านมาว่ามีตรงไหน ผิดพลาดอย่างไร ก่อนจะนำไปใช้จริง

**2.1.6 การนำไปใช้ (Deployment)** เพื่อให้การใช้ Model มีความถูกต้องบรรลุตามเป้าหมาย จะต้องมีการนำไปใช้จริง โดยมีขั้นตอนย่อยดังนี้

- Plan the Deployment: วางแผนสาหรับการติดตั้งใช้งาน
- Monitor and Maintain: ติดตามให้การช่วยเหลือระหว่างการใช้งาน
- Final Report: ทำรายงานสรุปผลการดำเนินการทั้งหมด

## 2.2 การเตรียมข้อมูล

การเตรียมข้อมูล (Data Preparation) เป็นการจัดเก็บข้อมูลและรวบรวมข้อมูลจากหลายแหล่ง ในบางครั้งข้อมูลอาจอยู่ในรูปแบบที่แตกต่างกันหรือมีความไม่สมบูรณ์ ซึ่งถ้าหากข้อมูลที่ได้มานั้นไม่มีคุณภาพจะส่งผลให้ผลลัพธ์ที่ได้มีคุณภาพต่ำไปด้วย ซึ่งเหตุผลสำคัญในการเตรียมข้อมูลสามารถแบ่งได้เป็นหัวข้อหลักๆ (Han and Kamber, 2000) ดังนี้ คือ 1) ข้อมูลไม่มีความสมบูรณ์ (Incomplete) เช่น มีข้อมูลบางตัวสูญหาย 2) ข้อมูลมีค่าสุดโต่ง (Outier) คือ ค่าข้อมูลที่จัดเก็บมากหรือน้อยเกินไปกว่าขอบเขตที่กำหนด 3) ข้อมูลไม่มีความสม่ำเสมอ (Inconsistent) คือ ข้อมูลอาจอยู่ในรูปแบบที่ต่างกันเช่นมีหน่วยที่ไม่สอดคล้องกัน และ 4) ข้อมูลที่ไม่ได้อยู่ในรูปแบบที่ประมวลผลได้ เช่น ข้อมูลอาจอยู่ในรูปแบบสัญลักษณ์หรือตัวอักษร เป็นต้น

### 2.2.1 การทำความสะอาดข้อมูล

การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนที่แก้ปัญหาค่าข้อมูลสูญหาย (Missing Value) โดยเราสามารถใช้วิธีจัดการค่าข้อมูลสูญหายได้ 3 วิธี (Roiger and Geatz, 2003) ดังต่อไปนี้

1) ถ้าตัวแปรที่เกิดค่าข้อมูลสูญหายมีค่าเป็นตัวเลขจำนวนจริง ทำการแทนค่าข้อมูลสูญหายด้วยค่าเฉลี่ยของข้อมูลรอบข้าง สามารถคำนวณได้จากสมการที่ (2.1)

$$\text{ค่าข้อมูลที่สูญหาย} = \frac{\text{ค่าก่อนหน้าข้อมูลสูญหาย} + \text{ค่าหลังค่าข้อมูลสูญหาย}}{2} \quad (2.1)$$

2) ถ้าตัวแปรที่เกิดค่าข้อมูลสูญหายมีค่าเป็นตัวเลขจำนวนนับจะแทนค่าข้อมูลสูญหายด้วยค่ามัธยฐานของกลุ่ม

3) ลบแถวข้อมูลที่มีค่าข้อมูลสูญหายทิ้ง

### 2.2.2 การเปลี่ยนรูปข้อมูล

การเปลี่ยนรูปข้อมูล (Data Transformation) เป็นขั้นตอนการแปลงข้อมูลที่ไม่พร้อมที่จะนำไปประมวลผลให้อยู่ในรูปที่พร้อมที่จะประมวลผล ซึ่งการเปลี่ยนรูปแบบข้อมูลสามารถแบ่งออกเป็น 3 วิธี (Roiger and Geatz, 2003) ดังต่อไปนี้

1) การเปลี่ยนข้อมูลรูปตัวเลข (Numerical Data) คือ การแปลงข้อมูลเข้าที่อยู่ในรูปตัวเลขให้อยู่ในช่วงค่าที่ต้องการ (Chantasut, 2004) แสดงได้ดังสมการ (2.2) เช่น การเปลี่ยนแปลงให้อยู่ในช่วง [0, 1] โดยค่าช่วงข้อมูลอยู่ระหว่าง 0 ถึง 1 เช่น ค่าข้อมูลเข้าคือ {100, 200, 300, 400} เมื่อผ่านการแปลงรูปข้อมูลจะได้ผลลัพธ์คือ {0.00, 0.33, 0.67, 1.0} เป็นต้น

$$\text{ค่าข้อมูลใหม่} = \frac{\text{ค่าข้อมูลเดิม} - \text{ค่าต่ำสุดช่วง}}{\text{ค่าสูงสุดช่วง} - \text{ค่าต่ำสุดช่วง}} \tag{2.2}$$

2) การเปลี่ยนรูปข้อมูลนามกำหนด (Nominal Data) คือ ใช้เทคนิคการแบ่งช่วงข้อมูลแทนค่าข้อมูลเป็นตัวเลข เช่น ข้อมูลทั้งหมด 4 สี คือ {สีแดง, สีเขียว, สีฟ้า, สีเหลือง} ถ้าต้องการแปลงข้อมูลให้อยู่ในช่วง [0, 1] จะแทนค่าข้อมูลได้ดังนี้ {0.00, 0.33, 0.67, 1.0}

3) การเปลี่ยนข้อมูลโดยการเพิ่มโหนดข้อมูลเข้า (Use of Additional Input Node) คือ มีการเพิ่มโหนด คือ สีแดง = [0, 0] สีเขียว = [0, 1] สีฟ้า = [1, 0] และสีเหลือง = [1, 1] เป็นต้น

## 2.3 การแบ่งช่วงข้อมูล

การแบ่งช่วงข้อมูล (Discretization) คือกระบวนในการแปลงคุณลักษณะของข้อมูลแบบต่อเนื่องให้อยู่ในรูปแบบของคุณลักษณะข้อมูลแบบไม่ต่อเนื่องเพื่อช่วยในการลดขนาดและความซับซ้อนของข้อมูล และเพิ่มประสิทธิภาพให้กับขั้นตอนวิธีของเหมืองข้อมูลประกอบด้วย วิธีการแบ่งช่วงข้อมูล คุณลักษณะของการแบ่งช่วงข้อมูล ขั้นตอนวิธีการแบ่งช่วงข้อมูล และเกณฑ์ที่ใช้ในการประเมินประสิทธิภาพ (Ming and Xinping, 2009) ซึ่งมีรายละเอียดดังต่อไปนี้

### 2.3.1 วิธีการแบ่งช่วงข้อมูล

วิธีการแบ่งช่วงข้อมูล (Discretization Methods) ประกอบด้วย 1) แบบแบ่งถัง 2) เอ็นโทรปี และ 3) หลักสถิติ อธิบายได้ดังต่อไปนี้

1) แบบแบ่งถัง (Binning) สำหรับการแบ่งช่วงข้อมูลแบบแบ่งถังนั้นเป็นวิธีการที่ง่ายที่สุดในการแบ่งช่วงข้อมูลเมื่อเปรียบเทียบกับวิธีการอื่นๆ โดยจะทำการเรียงลำดับ

ข้อมูล แล้วใช้หลักการแบ่งข้อมูลออกเป็นส่วนๆ แต่ละส่วนเรียกว่า Bin สำหรับตัวอย่างของการแบ่งช่วงข้อมูลแบบแบ่งถัง คือ 1) วิธีการแบ่งถังด้วยความกว้างที่เท่ากัน (Equal-Width) (Wong and Chiu, 1987) คือ การแบ่งช่วงข้อมูลออกเป็น n ช่วง ซึ่งแต่ละช่วงข้อมูลจะมีขนาดความกว้างของช่วงที่เท่ากันหมด (ใช้ระยะห่างเป็นเกณฑ์) สามารถคำนวณได้ดังสมการที่ (2.3) เช่น ชุดข้อมูลอายุมีจำนวนความกว้างของช่วงข้อมูลทั้งหมดคือ 1-90 ต้องการแบ่งออกเป็น 10 ถังข้อมูล ดังนั้นก็จะได้ขนาดของความกว้างของแต่ละช่วงข้อมูลคือ 9 และ 2) วิธีการแบ่งถังด้วยความถี่ที่เท่ากัน (Equal-Frequency) (Wong and Chiu, 1987) คือ แต่ละช่วงข้อมูลจะมีจำนวนข้อมูลที่เท่ากัน (ใช้ความถี่เป็นเกณฑ์) ซึ่งจะช่วยแก้ปัญหาเรื่องการกระจุกตัวของข้อมูลได้ สามารถคำนวณได้ดังสมการที่ (2.4) เช่น ชุดข้อมูลอายุมีจำนวนตัวอย่างข้อมูล 150 แถวข้อมูล ต้องการแบ่งออกเป็น 10 ถังข้อมูล ดังนั้นก็จะได้จำนวนความถี่ของข้อมูลในแต่ละถังคือ 15

$$\text{ความกว้างของแต่ละช่วง} = \frac{\text{ความกว้างของช่วงข้อมูลทั้งหมด}}{\text{จำนวนถังข้อมูล}} \tag{2.3}$$

$$\text{จำนวนข้อมูลแต่ละถัง} = \frac{\text{จำนวนข้อมูลทั้งหมด}}{\text{จำนวนถังข้อมูล}} \tag{2.4}$$

2) เอนโทรปีข้อมูล (Information Entropy) เป็นค่าที่ใช้วัดระดับความไม่เป็นระเบียบของข้อมูลว่ามีมากน้อยเพียงใด ถ้าข้อมูลมีความไม่เป็นระเบียบสูง ค่าเอนโทรปีก็จะยิ่งสูง แต่ถ้าข้อมูลมีความเป็นระเบียบสูง ค่าเอนโทรปีก็จะต่ำ ตัวอย่างการแบ่งช่วงข้อมูลที่ใช้ค่าเอนโทรปีข้อมูล เช่น MDLP (Fayyad and Irani, 1993) และ ID3 (Quinlan, 1993) โดยใช้ค่าเอนโทรปีข้อมูลที่ต่ำที่สุด (ข้อมูลมีความเป็นระเบียบมาก) เพื่อหาจุดตัดในการแบ่งแยกข้อมูลที่ดีที่สุด ซึ่งค่าเอนโทรปีข้อมูลสามารถคำนวณได้จากสมการที่ (2.5)

$$H(Y) = -\sum_{y \in Y} p(y)\log_2 p(y) \tag{2.5}$$

กำหนดให้ $H(Y)$ คือ ค่าเอนโทรปีของ Y

$p(y)$ คือ ความน่าจะเป็นของค่า y

3) หลักสถิติ (Statistical) เป็นวิธีการแบ่งช่วงข้อมูลที่ได้รับความนิยมอย่างมากในปัจจุบัน โดยอาศัยหลักการการหาค่าทางสถิติช่วยในการกำหนดจุดตัดที่ใช้ในการแบ่งช่วงข้อมูล หรือกำหนดจุดในการหลอมรวมกันของช่วงข้อมูลที่อยู่ติดกัน เช่น วิธีการของ ChiMerge จะทำการคำนวณหาค่า Chi-square ของช่วงข้อมูลที่อยู่ติดกันจากสมการที่ (2.6) แล้วทำการหลอมรวมของสองช่วงข้อมูลที่อยู่ติดกันที่มีค่า Chi-square น้อยที่สุดก่อน ทำไปเรื่อยจนกระทั่งค่า Chi-square ที่ได้จะเกินกว่าค่า Threshold ที่กำหนด (Kerber, 1992)

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{(A_{ij} - E_{ij})^2}{E_{ij}} \tag{2.6}$$

กำหนดให้ m คือ จำนวนของช่วงข้อมูลที่ใช้เปรียบเทียบ

k คือ จำนวนของกลุ่มเป้าหมาย

$A_{ij}$ คือ จำนวนตัวอย่างในช่วงข้อมูลที่ $i$ และในกลุ่มเป้าหมายของ j

$E_{ij}$ คือ ค่าความถี่ของ $A_{ij} = \frac{R_i \times C_j}{N}$ เมื่อ

$R_i$ คือ จำนวนของตัวอย่างในช่วงข้อมูลที่ $i$ ที่เท่ากับ $\sum_{j=1}^{k} A_{ij}$

$C_j$ คือ จำนวนของตัวอย่างในกลุ่มเป้าหมายที่ $j$ ที่เท่ากับ $\sum_{i=1}^{m} A_{ij}$

N คือ ผลรวมของตัวอย่าง ที่เท่ากับ $\sum_{j=1}^{k} C_i$

สำหรับตัวอย่างของขั้นตอนวิธีของการแบ่งช่วงข้อมูลโดยใช้หลักการทางสถิติ เช่น ChiMerge Chi2 และ CAIM เป็นต้น ซึ่งมีรายละเอียดดังต่อไปนี้

- Kerber (1992) นำเสนอขั้นตอนวิธีของ ChiMerge ซึ่งเป็น Supervised Discretization เป็นวิธีการที่ใช้ค่าทางสถิติ $\chi^2$ มาช่วยในการหาจำนวนช่วงข้อมูลที่ดีที่สุดในการแบ่งช่วงข้อมูล และใช้ค่า $\chi^2$-Threshold เป็นเกณฑ์ในการพิจารณาว่าจะมีการหลอมรวมช่วงข้อมูลที่อยู่ติดกันหรือไม่ วิธีนี้ผู้ใช้จำเป็นต้องกำหนดระดับค่า Significance Level ก่อนที่จะมีการแบ่งช่วงข้อมูล สำหรับขั้นตอนวิธีของ ChiMerge จะประกอบด้วย 2 ขั้นตอนดังนี้

1) เริ่มต้นด้วยการเรียงลำดับข้อมูล แล้วกำหนดขอบเขตของข้อมูลแต่ละช่วงข้อมูล โดยขอบเขตของแต่ละช่วงข้อมูลจะเป็นค่ากลางระหว่างค่าข้อมูลก่อนและค่าข้อมูลหลังของข้อมูลแต่ละตัวที่มีการเรียงลำดับแล้ว

2) ขั้นตอนการหลอมรวมกันของช่วงข้อมูล (Interval Merging) จะมีคุณลักษณะเป็น Bottom-up โดยมีขั้นตอนย่อย 2 ขั้นตอนดังนี้

- คำนวณค่า $\chi^2$ ดังสมการที่ (2.6) ของแต่ละช่วงข้อมูลที่อยู่กันติดกัน

- ทำการหลอมรวมกันของช่วงข้อมูลที่อยู่ติดกันที่มีค่า $\chi^2$ ที่น้อยที่สุดไปเรื่อยๆ จนกระทั่งทุกคู่ของช่วงข้อมูลมีค่า $\chi^2$ มากกว่าค่า $\chi^2$-Threshold ที่เรากำหนดไว้จึงจะหยุดการหลอมรวมกัน

- Liug และ Setiono (1995) นำเสนอขั้นตอนวิธี Chi2 ซึ่งเป็น Supervised Discretization ที่พัฒนาต่อจาก ChiMerge โดย Chi2 สามารถที่จะหาค่า $\chi^2$ Threshold ที่เหมาะสมได้โดยอัตโนมัติ เนื่องจากขั้นตอนวิธีของ ChiMerge ผู้ใช้จำเป็นต้องระบุค่า $\chi^2$-Threshold เพื่อเป็นเกณฑ์ในการพิจารณาว่าจะมีการหลอมรวมกันของช่วงข้อมูลที่อยู่ติดกันต่อไปอีกหรือไม่ จึงเป็นเรื่องยากที่ผู้ใช้จะสามารถระบุค่า $\chi^2$-Threshold ที่เหมาะสมได้ ดังนั้น Chi2 จึงแก้ปัญหาโดยใช้ Inconsistency Rate มาเป็นเกณฑ์ในการหยุดการหลอมรวมกันของช่วงข้อมูล แต่ข้อกำจัดของ Chi2 คือการคำนวณจะซับซ้อนกว่า ChiMerge สำหรับขั้นตอนวิธีในการแบ่งช่วงข้อมูลของ Chi2 จะประกอบด้วย 2 ขั้นตอนดังภาพประกอบ 2.2 และ 2.3

```
Phase 1:
    Set sigLevel = 0.5;
    Do while (InConsistency(data) < δ ) {
        for each numeric attribute {
            Sort(attribute, data);
            Chi-sq-initialization(attribute,data)
            do {
                chi-sq-calculation(attribute,data)
            } while (Merge(data))
        }
        sigLevel0 = sigLevel;
        sigLevel = decreSigLevel(sigLevel);
    }
```

ภาพประกอบ 2.2 ขั้นตอนวิธีของการแบ่งช่วงข้อมูลโดยใช้ Chi2 ในขั้นตอนที่ 1 (Liug และ Setiono, 1995)

```
Phase 2:
    Set all sigLvl[i] = sigLevel0 for attribute I;
    Do until no-attribute-can-be-merge {
        For each attribute I that can be merged {
            Sort(attribute, data);
            chi-sq-initialization(attribute,data);
            do {
                chi-sq-calculation(attribute, data)
            } while (Merge(data))
            If (InConsistency(data) < δ )
                sigLvl[i] = decreSigLevel(sigLvl[i]);
            else
                attribute I cannot be merged;
        }
    }
```

ภาพประกอบ 2.3 ขั้นตอนวิธีของการแบ่งช่วงข้อมูลโดยใช้ Chi2 ในขั้นตอนที่ 2 (Liug และ Setiono, 1995)

- Kurgan และ Cios (2004) นำเสนอขั้นตอนวิธีของ CAIM Discretization ซึ่งมีการประยุกต์ใช้ตาราง 2D Quanta Matrix แสดงดังตารางที่ 2.1 ซึ่งเป็นตารางกำหนดความถี่ของตัวแปรสองมิติคือ ตัวแปรแรกคือคลาสและตัวแปรที่สองเป็นค่าการแบ่งช่วงข้อมูลของคุณลักษณะ F

ตารางที่ 2.1 2D Quanta Matrix for Attribute F and Discretization Scheme D

| Class | Interval | | | | | Class Total |
|---|---|---|---|---|---|---|
| | $[d_0, d_1]$ | … | $(d_{r-1}, d_r]$ | … | $(d_{n-1}, d_n]$ | |
| $C_1$ | $q_{11}$ | … | $q_{1r}$ | … | $q_{1n}$ | $M_{1+}$ |
| : | : | ⋮ | : | ⋮ | : | : |
| $C_i$ | $q_{i1}$ | … | $q_{ir}$ | … | $q_{in}$ | $M_{i+}$ |
| : | : | ⋮ | : | ⋮ | : | : |
| $C_s$ | $q_{s1}$ | … | $q_{sr}$ | … | $q_{sn}$ | $M_{s+}$ |
| Interval Total | $M_{+1}$ | … | $M_{+r}$ | … | $M_{+n}$ | M |

กำหนดให้ $q_{ir}$ คือ ผลรวมของตัวอย่างข้อมูลที่เป็นของคลาส $C_i$ ที่อยู่ในช่วงข้อมูลของ $(d_{r-1}, d_r]$

$M_{i+}$ คือผลรวมของจำนวนตัวอย่างข้อมูลที่อยู่ในคลาส $C_i$

$M_{+r}$ คือผลรวมของจำนวนตัวอย่างที่อยู่ในช่วง $(d_{r-1}, d_r]$ ของคุณลักษณะ F

s คือจำนวนของคลาสทั้งหมด

n คือจำนวนของช่วงข้อมูล

โดยที่ i = 1, 2…, s และ r = 1, 2…, n

ขั้นตอนวิธีการแบ่งช่วงข้อมูล CAIM จะทดสอบทุกๆ จุดตัดที่เป็นไปได้ทั้งหมด และในแต่ละรอบของการทดสอบก็จะสร้างจุดตัดไปเรื่อยๆ จะทำการหยุดเมื่อเข้าตามเงื่อนไขให้หยุด แต่ละจุดตัดในแต่ละรอบสามารถคำนวณได้จากสมการที่ (2.7) โดยจะเลือกค่า CAIM ที่สูงที่สุดเป็นจุดตัด

$$\text{CAIM}\,(C, D\,|F) \;=\; \frac{\sum_{r=1}^{n} \frac{\max_r{}^2}{M_{+r}}}{n} \qquad (2.7)$$

กำหนดให้ C คือ คลาส

D คือ ตัวแปรช่วงข้อมูล ของคุณลักษณะ F

n คือ จำนวนของช่วงข้อมูลทั้งหมด

r คือ จำนวนรอบของช่วงข้อมูล

$\max_r$ คือ ค่าสูงสุดในคอลัมน์ที่ r

สำหรับขั้นตอนวิธีของ CAIM ผู้ใช้ไม่จำเป็นต้องมีการระบุพารามิเตอร์ในการแบ่งช่วงข้อมูล ประกอบด้วย 2 ขั้นตอน ขั้นตอนที่ 1 แสดงดังภาพประกอบ 2.4 คือ กำหนดขอบเขตของช่วงข้อมูล และขั้นตอนที่ 2 แสดงดังภาพประกอบ 2.5 กำหนดขอบเขตของช่วงข้อมูลใหม่ โดยใช้ค่าเกณฑ์ของ CAIM ที่สูงที่สุดในการกำหนดขอบเขตใหม่ของช่วงข้อมูลมีรายละเอียดดังต่อไปนี้

Given: Data consisting of M examples, S classes, and continuous attributes $F_i$

For every $F_i$ do:

Step 1.

1.1 Find maximum $(d_n)$ and minimum $(d_0)$ values of $F_i$

1.2 Form a set of all distinct values of $F_i$ in ascending order, and initialize all possible interval boundaries B with minimum, maximum and all the midpoints of all the adjacent pairs in the set

1.3 Set the initial discretization scheme as $D : \{[d_0, d_n]\}$, set GlobalCAIM = 0

1.4 Find maximum $(d_n)$ and minimum $(d_0)$ values of $F_i$

1.5 Form a set of all distinct values of $F_i$ in ascending order, and initialize all possible interval boundaries B with minimum, maximum and all the midpoints of all the adjacent pairs in the set

1.6 Set the initial discretization scheme as $D : \{[d_0, d_n]\}$, set GlobalCAIM = 0

ภาพประกอบ 2.4 ขั้นตอนวิธีของการแบ่งช่วงข้อมูลโดยใช้ CAIM ในขั้นตอนที่ 1

Step2.

2.1 Initialize k =1

2.2 Tentatively add and inner boundary, which is not already in D, from B, and calculate corresponding CAIM value

2.3 After all the tentative additions have been tried accept the one with the highest value of CAIM

2.4 If (CAIM > GlobalCAIM or k < S) then update D with the accepted in Step 2.3 boundary and set GlobalCAIM = CAIM, else terminate.

2.5 Set k= k+1 and go to 2.2

Output: Discretization scheme D

ภาพประกอบ 2.5 ขั้นตอนวิธีของการแบ่งช่วงข้อมูลโดยใช้ CAIM ในขั้นตอนที่ 2

**2.3.2 เกณฑ์ที่ใช้ในการประเมินประสิทธิภาพ**

เกณฑ์ที่ใช้ในการประเมินประสิทธิภาพ (Discretization Method Criterion) ขั้นตอนวิธีของการแบ่งช่วงข้อมูลจะใช้เพื่อเปรียบเทียบจุดแข็งจุดอ่อนของแต่ละขั้นตอนวิธีสามารถประเมินได้จากเกณฑ์ดังต่อไปนี้ คือ 1) ค่าความถูกต้อง 2) จำนวนของช่วงข้อมูล และ 3) เวลาที่ใช้ในการแบ่งช่วงข้อมูล (Garcia, J. Luengo, 2011)

1) ค่าความถูกต้อง (Accuracy Rate) เป็นประเมินจากค่าความถูกต้องในการเรียนรู้และการทดสอบของตัวจำแนก (Classifier) ที่ใช้ในทำนายผลและสร้างแบบจำลอง

2) จำนวนของช่วงข้อมูล (Number of Interval) ประเมินจากจำนวนของช่วงข้อมูลที่ได้หลังจากที่มีการแบ่งช่วงข้อมูลถ้าจำนวนของช่วงข้อมูลมีจำนวนมากก็จะส่งผลให้ใช้เวลาในการเรียนรู้มากขึ้น แต่ถ้าจำนวนของช่วงข้อมูลน้อยก็จะทำให้ใช้เวลาน้อยลงในการเรียนรู้

3) เวลาที่ใช้ (Time Require) เป็นประเมินประสิทธิภาพของขั้นตอนวิธีจากเวลาที่ใช้ในการแบ่งช่วงข้อมูล

## 2.4 การจำแนกประเภทข้อมูล

เทคนิคที่ใช้ในการจำแนกประเภทข้อมูล (Data Classification) ที่ใช้ในงานวิจัยนี้ ประกอบด้วย ต้นไม้ตัดสินใจ คือ J48 โครงข่ายประสาทเทียม คือ Radial Basis Function (RBF) Multilayer Perceptron (MLP) และนาอีฟเบย์ (Naïve Bays)

### 2.4.1 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree) เป็นการจำแนกข้อมูลโดยแทนความรู้ในรูปแบบของต้นไม้โดยที่แต่ละโหนด (Node) แสดงคุณลักษณะที่ใช้ในการทดสอบข้อมูล แต่ละกิ่ง (Branch) แสดงผลในการทดสอบและโหนดใบ (Leaf Node) แสดงแสดงกลุ่มหรือคลาสที่กำหนดไว้ (กาญจนา, 2551) ขั้นตอนวิธีการเรียนรู้ของต้นไม้ตัดสินใจมีดังนี้

1) ค้นหาคุณลักษณะข้อมูล (Attribute) ที่สำคัญที่สุดในการแบ่งข้อมูล โดยคุณลักษณะข้อมูลนี้จะถูกนำมาสร้างเป็นโหนดราก โดยมีโหนดใบ (Leaf Node) เป็นผลลัพธ์ที่ถูกกำหนดไว้ก่อน ซึ่งในการเลือกคุณลักษณะข้อมูลมาเป็นโหนดจะใช้เกณฑ์สารสนเทศ มาพิจารณาในการเลือกโหนดแต่ละโหนดของต้นไม้ ถ้าคุณลักษณะข้อมูลใดที่มีค่าเกณฑ์สารสนเทศสูงสุดหรือว่ามีค่าเอ็นโทรปีน้อยสุด ก็จะถูกเลือกให้เป็นโหนดราก

2) นำค่าที่เป็นไปได้ในคุณลักษณะข้อมูลที่ถูกเลือกมาแบ่งแยก (Split) เป็นกิ่งออกจากโหนดที่ได้เลือกไว้

3) วนกลับไปทำในขั้นตอนแรก คือหาคุณลักษณะข้อมูลที่สำคัญที่สุดจาก จากข้อมูลที่เข้าเพื่อหาตัวแบ่งแยกของโหนดถัดไป

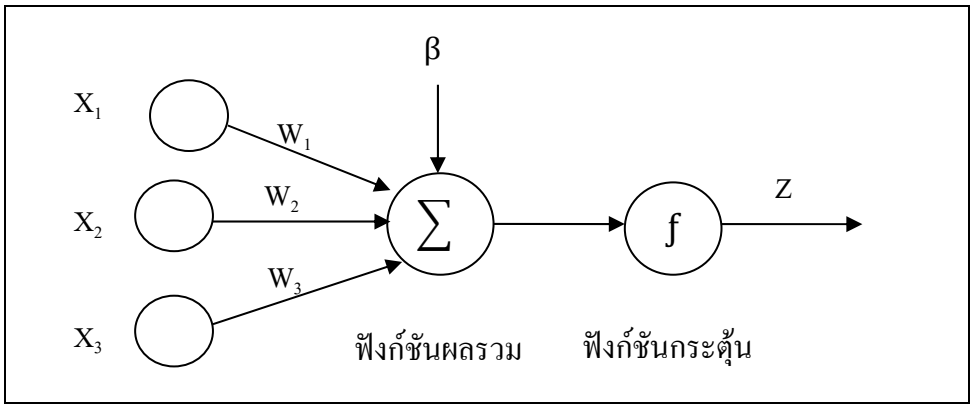
4) ทำซ้ำไปเรื่อยๆ จนกระทั่งตัวอย่างข้อมูล (Instance) ทุกตัวมีค่าคลาสเหมือนกันหมดหรือไม่มีคุณลักษณะข้อมูลใดเหลือในการแบ่งแยก

### 2.4.2 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นรูปแบบการประมวลผลที่เลียนแบบการทำงานของสมองมนุษย์ ประกอบด้วยหน่วยประมวลผลย่อยหรือเพอร์เซฟตรอน (Perceptron) หลายหน่วยเชื่อมเชื่อมต่อกันเป็นโครงข่าย ข้อดีของโครงข่ายประสาทเทียมคือ สามารถทำนายค่าที่มีความแม่นยำสูง ทนทานต่อความผิดพลาด และสามารถรองรับข้อมูลที่ไม่สมบูรณ์หรือมีสิ่งรบกวน (Mellit and Kalogirou, 2008; Shiva and Khare, 2004)

**2.4.2.1 เพอร์เซฟตรอน**

เพอร์เซฟตรอน (Perceptron) เป็นหน่วยประมวลผลที่เล็กที่สุดของโครงข่ายประสาทเทียม องค์ประกอบของเพอร์เซฟตรอนประกอบด้วย ฟังก์ชันผลรวม (Summation Function) ทำหน้าที่หาผลรวมของผลคูณระหว่างค่านำหนักของข้อมูลเข้ากับค่าของข้อมูลเข้า และฟังก์ชันกระตุ้น (Activation Function) ทำหน้าที่แปลงผลลัพธ์จากฟังก์ชันผลรวมให้อยู่ในช่วงค่าที่ต้องการ แสดงภาพองค์ประกอบของเพอร์เซฟตรอนได้ดังภาพประกอบ 2.6




ภาพประกอบ 2.6 องค์ประกอบของเพอร์เซฟตรอน

การคำนวณของฟังก์ชันผลรวมสามารถแสดงได้ดังสมการที่ (2.8) ดังนี้

$$y = \sum_{i=1}^{n} w_i x_i + \beta \qquad (2.8)$$

กำหนดให้ $y$ คือ ผลลัพธ์ของฟังก์ชันผลรวม

$x_i$ คือ ค่าข้อมูลเข้าตัวที่ i

$n$ คือ จำนวนข้อมูลเข้าทั้งหมด

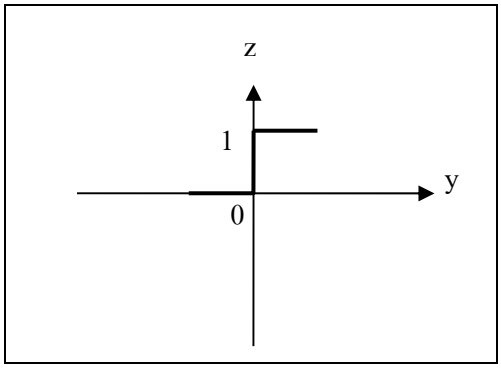
$w_i$ คือ ค่าน้ำหนักของข้อมูลเข้าตัวที่ i

$\beta$ คือ ค่าโน้มถ่วงเอียง

ฟังก์ชันกระตุ้น (f) ทำหน้าที่แปลงผลลัพธ์ของฟังก์ชันผลรวมให้อยู่ในช่วงค่าที่ต้องการ ซึ่งมีหลายแบบ ตัวอย่างของฟังก์ชันกระตุ้นสามารถแสดงได้ ดังสมการที่ (2.10) ถึง (2.13) โดยกำหนดให้ z คือ ผลลัพธ์ของฟังก์ชันกระตุ้น และ y คือ ผลลัพธ์ของฟังก์ชันผลรวม

1) ฟังก์ชันสเตป (Step Function) ผลลัพธ์ที่ได้จะเป็นค่า 0 และ 1 แสดงดังสมการที่ (2.9) และภาพประกอบ 2.7

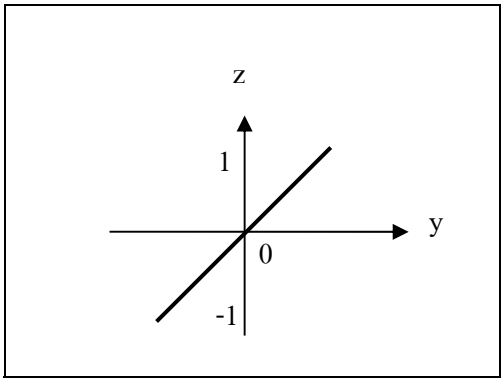
$$Z = \begin{cases} 1 & ; \text{if } y \geq 0 \\ 0 & ; \text{otherwise} \end{cases} \qquad (2.9)$$




ภาพประกอบ 2.7 ฟังก์ชันสเตป

2) ฟังก์ชันลิเนีย (Linear Function) ผลลัพธ์ที่ได้จะมีค่าเท่ากับข้อมูลเข้า แสดงดังสมการที่ (2.10) และภาพประกอบ 2.8

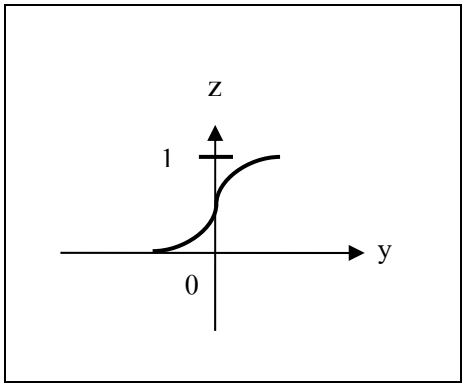
$$z = y \qquad (2.10)$$




ภาพประกอบ 2.8 ฟังก์ชันลิเนีย

3) ฟังก์ชันล๊อกซิกมอยด์ (Log-Sigmoid Function) ผลลัพธ์ที่ได้อยู่ใช่ช่วง 0 ถึง 1 แสดงดังสมการที่ (2.11) และภาพประกอบ 2.9

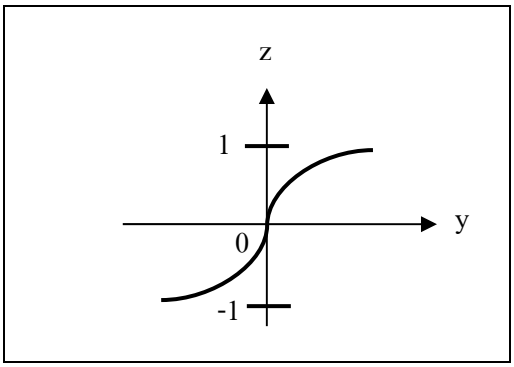
$$z = \frac{1}{1+e^{-y}} \tag{2.11}$$




ภาพประกอบ 2.9 ฟังก์ชันล็อกซิกมอยด์

4) ฟังก์ชันแทนซิกมอยด์ (Tan-Sigmoid Function) ผลลัพธ์ที่ได้จะอยู่ในช่วง -1 ถึง 1 แสดงดังสมการที่ (2.12) และภาพประกอบ 2.10

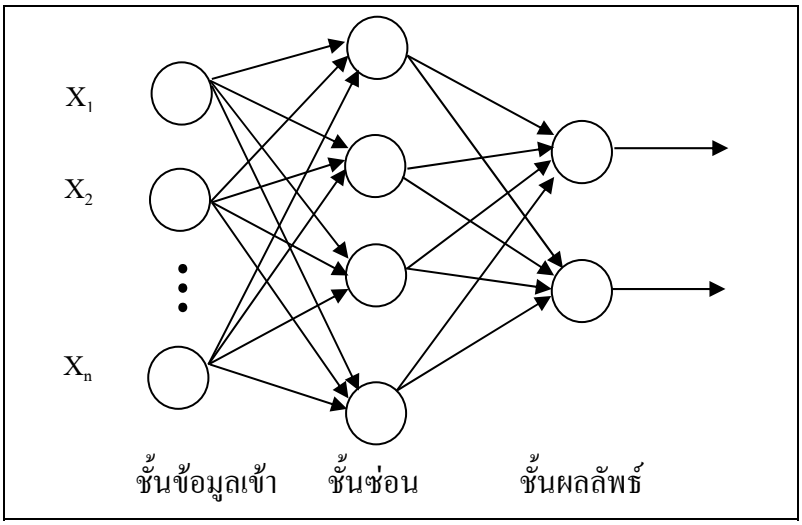
$$z = \frac{2}{1+e^{-2y}} - 1 \tag{2.12}$$




ภาพประกอบ 2.10 ฟังก์ชันแทนซิกมอยด์

**2.4.2.2 โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น**

โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron Neural Network) เป็นการนำเอาเพอร์เซฟตรอนหลายหน่วยมาเชื่อมต่อกันเป็นโครงข่าย เพื่อเพิ่มประสิทธิภาพในการพยากรณ์ รูปแบบการเรียนรู้ของโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นจะเป็นการเรียนรู้แบบผู้สอน จะต้องมีการสอนโครงข่ายประสาทเทียมก่อนนำไปใช้งานจริง โครงสร้างของโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นมี 3 ระดับ คือ ชั้นข้อมูลเข้า ชั้นซ่อน และชั้นผลลัพธ์ โดยชั้นข้อมูลเข้ามี 1 ชั้น ชั้นซ่อนมีกี่ชั้นก็ได้ และชั้นผลลัพธ์มี 1 ชั้น ในแต่ละชั้นจะมีเพอร์เซฟตรอนกี่หน่วยก็ได้แสดงดังภาพประกอบ 2.11




ภาพประกอบ 2.11 โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า (Feed Forward Networks)

**2.4.3.3 เรดียลเบสิสฟังก์ชัน**

เรดียลเบสิสฟังก์ชัน หรือ RBFNetwork (Radial Basis Function Neural Network) เป็นโครงข่ายประสาทเทียมประกอบด้วย 3 Layer คือ Input Layer Hidden Layer และ Output Layer (Lan and Frank, 2005) โดย Hidden Unit มีรูปแบบการประมวลผลโดยใช้ฟังก์ชันกระตุ้นแบบเรเดียล (Radail Activated Function) (Nikolaev, 2008) ซึ่งแสดงดังสมการที่ (2.13) ถึง (2.15)

1) Mutiquadratics: $\varphi(x) = (x^2 + c^2)^{\frac{1}{2}}$ เมื่อ $c < 0$ (2.13)

2) Inverse multiquadratics: $\varphi(x) = 1/(x^2 + c^2)^{\frac{1}{2}}$ เมื่อ $c > 0$ (2.14)

3) Gaussian: $\varphi(x) = \exp(-x^2/2\sigma^2)$ เมื่อ $\sigma > 0$ (2.15)

โดยทั่วไปจะใช้ Gaussian Function เป็น Radial Activated Function โดยผลลัพธ์ของฟังก์ชันกระตุ้นแบบเรเดียลจะอยู่ในช่วง (0, 1) ดังสมการที่ (2.16)

$$F(x) = \sum_{i=1}^{n} w_i \exp(-\|x - x_i\|^2/2\sigma_i^2) \quad (2.16)$$

เมื่อ $w_i$ คือ เป็นน้ำหนักของเอาร์พุทระหว่าง Hidden Unit และ Output Unit

n คือ จำนวน basis function

$x_i$ คือ ศูนย์กลางของ basis function

x คือ ค่าข้อมูลเข้า

**2.4.3 นาอีฟเบย์**

นาอีฟเบย์ (Naïve Bayes) เป็นเทคนิคการจำแนกข้อมูลพัฒนาโดย Thomas Bayes โดยมีการตั้งสมมติฐานเพื่อกำหนดให้การเกิดของเหตุการณ์ต่างๆ ที่ใช้ในการจัดกลุ่มนั้นเป็นอิสระต่อกัน ซึ่งจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละตัวกับตัวแปรตามเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นของแต่ละความสัมพันธ์จุดประสงค์เพื่อต้องการสร้างแบบจำลองที่อยู่ในรูปของความน่าจะเป็น เพื่อหาว่าสมมติฐานใดถูกต้องมีที่สุด ข้อดีของวิธีการแบบเบย์คือสามารถใช้ข้อมูลและความรู้ก่อนหน้านี้ เพื่อใช้ในการเรียนรู้ นอกจากนี้ยังเหมาะสมกับชุดข้อมูลที่มีขนาดใหญ่และคุณลักษณะข้อมูลที่เป็นอิสระต่อกัน ทฤษฎีเบย์ (Bayes' Theorem) (John, Pat Langley, 1995) สามารถคำนวณได้ดังสมการที่ (2.17)

(2.17)

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

กำหนดให้ P(H) คือ ความน่าจะเป็นที่จะเกิดสมมติฐาน H

P(E) คือ ความน่าจะเป็นของชุดข้อมูล E

P(H|E) คือ ความน่าจะเป็นของ H เมื่อทราบ E

P(E|H) คือ ความน่าจะเป็นของ E เมื่อทราบ H

**2.4.4 การประเมินประสิทธิภาพตัวจำแนกประเภทข้อมูล**

การประเมินตัวจำแนกประเภท (Classifier Evaluation) สามารถแบ่งออกเป็น 3 ด้านดังนี้คือ ด้านประสิทธิภาพในการสอนระบบ (Training Efficiency) ด้านประสิทธิภาพในการจำแนก (Classification Efficiency) และความถูกต้องในการจำแนก (Correctness of Classification) ซึ่งประสิทธิภาพในการสอนระบบและการจำแนกประเภทจะวัดจากความเร็วในการประมวลผลและการใช้เนื้อที่หน่วยความจำเป็นหลัก (Radovanovic, 2006) การวัดประสิทธิภาพนิยมใช้วิธีทางด้านการค้นคืนสารสนเทศ ซึ่งการวัดประสิทธิภาพของระบบจะประเมินจากข้อมูลที่ได้จากการจำแนกประเภทแสดงเป็นตาราง Confusion Matrix (Kohavi and Provost, 1998) ประกอบด้วยค่าข้อมูลที่เป็นค่าจริง (Actual) และข้อมูลที่เป็นค่าทำนาย (Predicted) ดังตารางที่ 2.2 แสดง Confution Maxtrix สำหรับการจำแนกประเภทที่มี 2 คลาส

ตารางที่ 2.2 Confusion Maxtrix

| Value | | Predicted | |
|---|---|---|---|
| | | Negative | Positive |
| Actual | Negative | a | b |
| | Positive | c | d |

กำหนดให้ a คือ จำนวนตัวอย่างที่ทำนายได้ถูกต้องว่าตัวอย่างเป็น Negative

b คือ จำนวนตัวอย่างที่ทำนายผิดว่าตัวอย่างเป็น Positive

c คือ จำนวนตัวอย่างที่ทำนายผิดว่าตัวอย่างเป็น Negative

d คือ จำนวนตัวอย่างที่ทำนายได้ถูกต้องว่าตัวอย่างเป็น Positive

จาก Confusion Maxtrix ในตารางที่ 2.2 สามารถคำนวณค่าต่างๆ ได้ดังนี้

1) Accuracy (AC) เป็นสัดส่วนของจำนวนทั้งหมดที่ตัวจำแนกประเภททำนายได้ถูกต้อง สามารถคำนวณได้ดังสมการที่ (2.18)

$$AC = \frac{a+d}{a+b+c+d} \tag{2.18}$$

2) Recall (R) หรือ Time Positive (TP) เป็นสัดส่วนของจำนวนตัวอย่างที่ทำนายได้ถูกต้อง กรณีที่ค่าจริงเป็น Positive สามารถคำนวณได้ดังสมการที่ (2.19)

$$R = \frac{d}{c+d} \tag{2.19}$$

2) Precision (P) เป็นสัดส่วนของจำนวนตัวอย่างที่ทำนายได้ถูกต้อง กรณีมีค่าทำนายเป็น Positive สามารถคำนวณได้ดังสมการที่ (2.20)

$$P = \frac{d}{b+d} \tag{2.20}$$

**2.4.5 การแบ่งช่วงข้อมูลในการทดสอบ**

เป็นวิธีการในการคาดการณ์โมเดล โดยการทำงานของ K-Folds Cross Validation เป็นการแบ่งข้อมูลออกเป็น K ชุดๆเท่ากัน ในการทำงานเป็นชุดสอน (Train Set) และชุดทดสอบ (Test Set) โดยทำงานทั้งหมด K ครั้ง การทำงานรอบแรกข้อมูลชุดที่ 1 จะเป็นชุดทดสอบข้อมูลชุดที่เหลือ (K-1) จะเป็นชุดสอน และในรอบต่อไปข้อมูลชุดที่ 2 จะเป็นชุดทดสอบ ข้อมูลชุดที่เหลือจะเป็นชุดสอนจนครบทั้งหมด K รอบ ข้อดีของวิธีนี้คือ ข้อมูลทุกตัวจะมีโอกาสเป็นทั้งชุดสอนและชุดทดสอบ และในการสอนแต่ละครั้งจะมีข้อมูลจากทุกคลาส การเลือกจำนวน Fold จะพิจารณาจากจำนวนตัวอย่าง หากจำนวนตัวอย่างมีจำนวนมากสามารถเลือกจำนวน Fold ที่เหมาะสมได้ดี ตัวอย่างการทำงาน 10-Folds Cross Validation แสดงดังตารางที่ 2.3

ตารางที่ 2.3 ตัวอย่างการทำงานของ K-Folds Cross Validation

| รอบที่ | ชุดข้อมูล K-Fold | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1 | ■ | | | | | | | | | |
| 2 | | ■ | | | | | | | | |
| 3 | | | ■ | | | | | | | |
| 4 | | | | ■ | | | | | | |
| 5 | | | | | ■ | | | | | |
| 6 | | | | | | ■ | | | | |
| 7 | | | | | | | ■ | | | |
| 8 | | | | | | | | ■ | | |
| 9 | | | | | | | | | ■ | |
| 10 | | | | | | | | | | ■ |

■ ข้อมูลชุดทดสอบ

□ ข้อมูลชุดสอน

# บทที่ 3

# การออกแบบขั้นตอนวิธีในการแบ่งช่วงข้อมูลสำหรับเหมืองข้อมูล

วิทยานิพนธ์นี้มุ้งเน้นไปที่การออกแบบและพัฒนาขั้นตอนวิธีในการแบ่งช่วงข้อมูลสำหรับการทำเหมืองข้อมูลเพื่อให้มีประสิทธิภาพ โดยใช้หลักการทางสถิติมาช่วยในการพิจารณาค่าที่เหมาะสมที่สุดในการที่จะหลอมรวมกันของสองช่วงข้อมูลที่อยู่ติดกัน โดยมีการประยุกต์ใช้ตาราง 2D Quanta matrix ในการหาความสัมพันธ์ของการกระจายตัวระหว่างคลาสและแอตทริบิวต์แล้วใช้ค่าเฉลี่ยการกระจายตัวของข้อมูลในทุกๆ ช่วงข้อมูลเป็นเกณฑ์ในการหลอมรวมกันของสองช่วงข้อมูลที่อยู่ติดกันเพื่อหาช่วงข้อมูลที่ดีที่สุดในการแบ่งช่วงข้อมูล

การออกแบบขั้นตอนวิธีในการแบ่งช่วงข้อมูลสำหรับเหมืองข้อมูลที่มีประสิทธิภาพสามารถแบ่งออกเป็น 3 ขั้นตอนหลักคือ 1) ขั้นตอนของการเตรียมข้อมูล 2) ขั้นตอนของการแบ่งช่วงข้อมูล และ 3) ขั้นตอนในการทดสอบประสิทธิภาพในการทำงานของขั้นตอนวิธีที่นำเสนอ โดยมีรายละเอียดดังต่อไปนี้

## 3.1 ขั้นตอนของการเตรียมข้อมูล

การเตรียมข้อมูล (Data Preparation) มีจุดประสงค์เพื่อเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมจะประมวลผลข้อมูล เนื่องจากการจัดเก็บอาจมีการรวบรวมข้อมูลจากหลายๆ แหล่ง ในบางครั้งข้อมูลอาจจะอยู่ในรูปแบบที่แตกต่างกันหรือมีความไม่สมบูรณ์ จึงจำเป็นต้องมีการเตรียมข้อมูลก่อนการทำเหมืองข้อมูล ซึ่งวิธีการในการเตรียมข้อมูลนั้นมีหลายวิธีการขึ้นอยู่กับความต้องการของผู้ใช้ที่จะนำไปใช้งาน เช่น การแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยของข้อมูลรอบข้าง ถ้าชุดข้อมูลนั้นมีการสูญหายในแถวที่อยู่ติดกันไม่เกินหนึ่งแถวข้อมูลหรือมีข้อมูลสูญหายไม่มากหนัก หรือวิธีการหาค่าเฉลี่ยของข้อมูลทั้งหมดมาแทนค่าข้อมูลที่สูญหาย ถ้าชุดข้อมูลนั้นมีการสูญหายในแถวที่อยู่ติดกันมากกว่าหนึ่งแถว หรือวิธีการเปลี่ยนรูปตัวแปรข้อมูลเข้าให้อยู่ในช่วง [0, 1] และวิธีการเปลี่ยนรูปข้อมูลผลลัพธ์ให้อยู่ในรูปของ 0 และ 1 สำหรับวิทยานิพนธ์นี้ผู้วิจัยได้เลือกใช้วิธีการของการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้าง เนื่องจากการแบ่งช่วงข้อมูลเป็นการแปลงข้อมูลจากชุดข้อมูลแบบต่อเนื่องให้เป็นแบบไม่ต่อเนื่องและชุดข้อมูลที่ใช้มีค่าข้อมูลสูญหาย

เพียงเล็กน้อย ดังนั้นควรเลือกวิธีการทำความสะอาดข้อมูลโดยใช้วิธีการของการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้าง โดยมีขั้นตอนการทำงานดังต่อไปนี้

การแทนค่าข้อมูลที่สูญหาย (Missing Value) ด้วยค่าเฉลี่ยของข้อมูลรอบข้าง สามารถคำนวณด้วยสมการที่ 2.1 ในวิทยานิพนธ์เล่มนี้ ผู้วิจัยได้นำตัวอย่างชุดข้อมูลอายุแสดงดังตารางที่ 3.1 มาใช้เพื่อช่วยในการอธิบายในแต่ละขั้นตอนเพื่อให้สามารถอธิบายและเข้าใจแต่ละขั้นตอนได้ง่ายขึ้น ตัวอย่างชุดข้อมูลอายุประกอบด้วย 30 แถวข้อมูล และมี 3 คลาส โดยมีค่าข้อมูลสูญหาย 2 ค่า คือ ค่าข้อมูลของตัวแปรอายุ (age) ลำดับที่ (id) 5 และ 23

ตารางที่ 3.1 ตัวอย่างข้อมูลดิบของชุดข้อมูลอายุ

| id | age | class | | id | age | class |
|---|---|---|---|---|---|---|
| 1 | 1 | care | | 16 | 11 | edu |
| 2 | 1 | care | | 17 | 11 | edu |
| 3 | 2 | care | | 18 | 12 | work |
| 4 | 2 | care | | 19 | 15 | edu |
| 5 | - | care | | 20 | 16 | edu |
| 6 | 2 | care | | 21 | 16 | edu |
| 7 | 2 | care | | 22 | 19 | edu |
| 8 | 3 | care | | 23 | - | edu |
| 9 | 4 | edu | | 24 | 25 | work |
| 10 | 4 | edu | | 25 | 27 | work |
| 11 | 5 | edu | | 26 | 28 | work |
| 12 | 7 | edu | | 27 | 28 | work |
| 13 | 7 | edu | | 28 | 31 | edu |
| 14 | 9 | care | | 29 | 33 | work |
| 15 | 10 | edu | | 30 | 35 | work |

การเตรียมข้อมูลด้วยวิธีการแทนค่าข้อมูลที่สูญหายของข้อมูลลำดับที่ 5 จากชุดข้อมูลอายุ ค่าที่จะใช้ในการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้างคือ ค่าข้อมูลลำดับที่ 4 และค่าข้อมูลที่อยู่ในลำดับที่ 6 จะได้คือ $(2 + 2) / 2 = 2$ ดังนั้นก็จะแทนค่าข้อมูลที่สูญหายในลำดับข้อมูลที่ 5 คือ 2 และสำหรับข้อมูลลำดับที่ 23 ค่าที่ใช้ในการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้างคือ ค่าข้อมูลลำดับที่ 22 และค่าข้อมูลที่อยู่ในลำดับที่ 24 จะได้คือ $(19 + 25) / 2 = 22$ ดังนั้นก็จะแทนค่าข้อมูลที่สูญหายในลำดับข้อมูลที่ 23 คือ 22 แสดงดังตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างข้อมูลดิบของชุดข้อมูลอายุที่ผ่านการแทนค่าข้อมูลที่สูญหาย

| id | age | class | | id | age | class |
|---|---|---|---|---|---|---|
| 1 | 1 | care | | 16 | 11 | edu |
| 2 | 1 | care | | 17 | 11 | edu |
| 3 | 2 | care | | 18 | 12 | work |
| 4 | 2 | care | | 19 | 15 | edu |
| 5 | **2** | care | | 20 | 16 | edu |
| 6 | 2 | care | | 21 | 16 | edu |
| 7 | 2 | care | | 22 | 19 | edu |
| 8 | 3 | care | | 23 | **22** | edu |
| 9 | 4 | edu | | 24 | 25 | work |
| 10 | 4 | edu | | 25 | 27 | work |
| 11 | 5 | edu | | 26 | 28 | work |
| 12 | 7 | edu | | 27 | 28 | work |
| 13 | 7 | edu | | 28 | 31 | edu |
| 14 | 9 | care | | 29 | 33 | work |
| 15 | 10 | edu | | 30 | 35 | work |

## 3.2 ขั้นตอนวิธีของการแบ่งช่วงข้อมูล

ขั้นตอนวิธีของการแบ่งช่วงข้อมูลโดยใช้ค่าเฉลี่ยการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ของทุกๆ ช่วงข้อมูล หรือเรียกว่า Class Attribute Interval Average Discretization Algorithm (CAIA) โดยมีการประยุกต์ใช้หลักการทางสถิติและตาราง 2D quanta matrix ดังแสดงในตารางที่ 3.3 มาช่วยในการหาช่วงข้อมูลที่ดีที่สุด CAIA สามารถแบ่งออกเป็น 2 ขั้นตอนหลักๆ คือ ขั้นตอนที่ 1 การเรียงลำดับข้อมูลเพื่อหาขอบเขตที่จะเริ่มต้นในการแบ่งช่วงข้อมูลและจัดการข้อมูลให้อยู่ในรูปแบบของ 2D quanta matrix แสดงดังภาพประกอบ 3.1 ซึ่งเป็นขั้นตอนวิธีที่ใช้ในการแบ่งช่วงข้อมูลของ CAIA และส่วนขั้นตอนที่ 2 เป็นการค้นหาช่วงข้อมูลที่ดีที่สุดที่จะใช้ในการหลวมรวมของสองช่วงข้อมูลที่อยู่ติดกันโดยใช้ค่าเฉลี่ยของการกระจายตัวระหว่างคลาสและแอตทริบิวต์ในทุกๆ ช่วงข้อมูลแสดงดังภาพประกอบ 3.2

ตารางที่ 3.3 2D Quanta matrix for attribute A and discretization Scheme D

| Class | Interval | | | | | Total |
|---|---|---|---|---|---|---|
| | 1 | … | r | | n | |
| | $[d_1,d_2]$ | … | $(d_{r-1},d_r]$ | … | $(d_{n-1},d_n]$ | |
| $C_1$ | $q_{11}$ | … | $q_{1r}$ | … | $q_{1n}$ | $M_{1+}$ |
| $C_i$ | $q_{i1}$ | … | $q_{ir}$ | … | $q_{in}$ | $M_{i+}$ |
| $C_s$ | $q_{s1}$ | … | $q_{sr}$ | … | $q_{sn}$ | $M_{s+}$ |
| Interval Total | $M_{+1}$ | … | $M_{+r}$ | … | $M_{+n}$ | M |
| $CAI_{+r}$ | $CAI_{+1}$ | | $CAI_{+r}$ | | $CAI_{+n}$ | CAIA |

กำหนดให้ s คือจำนวนของคลาส

n คือจำนวนของช่วงข้อมูล

M คือจำนวนของ instances

$q_{ir}$ คือผลรวมของข้อมูลจากคลาสที่ i ที่อยู่ในช่วงข้อมูลของ $(d_{r-1},d_r]$

$M_{i+}$ คือผลรวมของข้อมูลทั้งหมดที่อยู่ในคลาสที่ i ของทุกช่วงข้อมูล

$M_{+r}$ คือผลรวมของข้อมูลทั้งหมดที่อยู่ในช่วงข้อมูลของ $(d_{r-1},d_r]$ ของทุกคลาส

$CAI_{+r}$ คือค่าการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ในทุกๆ ช่วงข้อมูลที่ r

CAIA คือค่าเฉลี่ยการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ในทุกๆ ช่วงข้อมูล

โดยที่ i มีค่าตั้งแต่ 1 ถึง s และ r มีค่าตั้งแต่ 1 ถึง n

Given: Data set where M is the total number of instances, s is the total number of classes, and x is the total number of attributes

Step1 . Create 2D-quanta matrix

1.1 For p = 1 to x (each continuous attribute $A_p$)

1.2 Let $d_1$ = minimum values of $A_p$

1.3 Let $d_n$ = maximum value of $A_p$

1.4 Sort all distinct values of $A_p$ in ascending order

1.5 Merge all adjacent intervals that belongs to the same class.

1.6 For r = 1 to n // n is the number of intervals

1.7 //Calculate midpoint of each adjacent intervals

$$B_{+r} = (Bmax_r + Bmin_{r+1})/2$$

1.8 Output is the 2D-quanta matrix, discretization scheme D for attribute $A_p$

ภาพประกอบ 3.1 ขั้นตอนการสร้าง 2D quanta matrix ของการแบ่งช่วงข้อมูลโดยใช้ CAIA

Step2 . Merge using class attribute interval average algorithm

2.1 For r = 1 to n // n is the total number of intervals

2.2 For i = 1 to s

2.3 // Calculate Class Attribute Interval

$$CAI_{+r} = \sum_{i=1}^{s} (q_{ir}{}^2 / M_{i+}) \times (M_{+r})$$

2.4 If (n > s) then

2.5 $CAI_{min}$ = $CAI_{+r}$ minimum value

2.6 candidate merge $CAI_{min}$ with $CAI_{right\_of\_min}$,

2.7 $$CAIA_{right} = \sum_{r=1}^{n} CAI_{+r}/n$$

2.8 candidate merge $CAI_{min}$ with $CAI_{letf\_of\_min}$,

2.9 $$CAIA_{letf} = \sum_{r=1}^{n} CAI_{+r}/n$$

2.10 If ($CAIA_{left} > CAIA_{right}$)

2.11 then $CAIA = CAIA_{left}$

2.12 else $CAIA = CAIA_{right}$

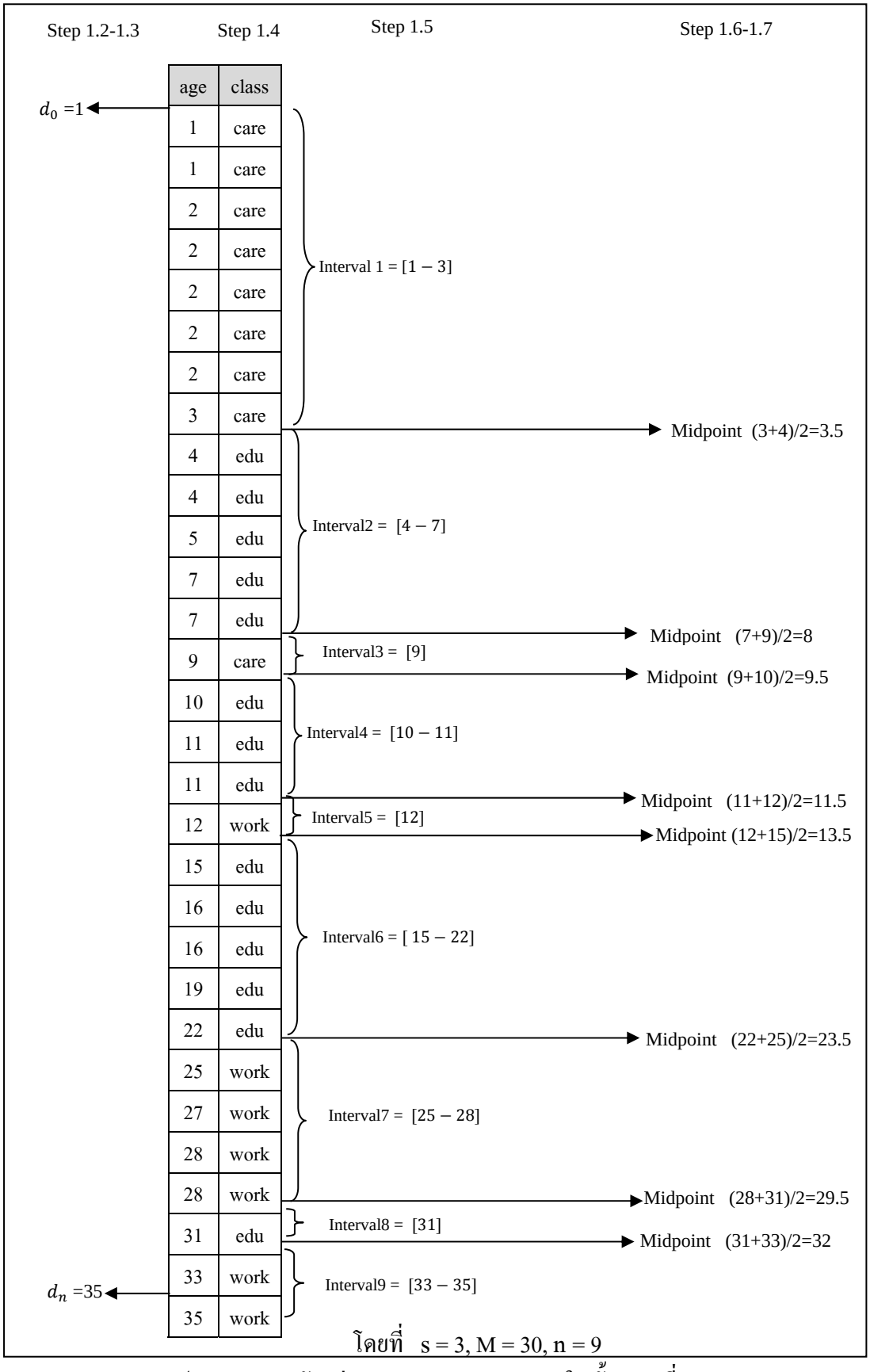
2.13 n = n – 1

2.14 Output discretization scheme D' for attribute $A_p$

ภาพประกอบ 3.2 ขั้นตอนการหลอมรวมของการแบ่งช่วงข้อมูลโดยใช้ CAIA

ขั้นตอนวิธีในการแบ่งช่วงข้อมูลของ CAIA ประกอบด้วยขั้นตอนย่อยโดยสามารถอธิบายได้ดังต่อไปนี้

**ขั้นตอนที่ 1** เป็นการจัดรูปแบบข้อมูลให้อยู่ในรูปของตาราง 2D quanta matrix เพื่อแสดงความสัมพันธ์ระหว่างค่าคลาส $C_1$ จนถึง $C_s$ กับ n ช่วงข้อมูล ที่มีจำนวนข้อมูลเป็น M instances ของแอตทริบิวต์ $A_p$ ที่ p = 1 to x เมื่อ x คือจำนวนของแอตทริบิวต์ สำหรับตัวอย่างข้อมูลในภาพประกอบ 3.3 เป็นตัวอย่างชุดข้อมูลอายุที่ได้ผ่านขั้นตอนของการเตรียมข้อมูลเรียบร้อย ประกอบด้วยจำนวนข้อมูลทั้งหมด 30 instances และมีจำนวนคลาส 3 คลาส คือ care,




ภาพประกอบ 3.3 ตัวอย่างการทำงานของ CAIA ในขั้นตอนที่ 1.1-1.7

edu (Education) และ work และขั้นตอนวิธีในการแบ่งช่วงข้อมูลของ CAIA ในขั้นตอนที่ 1 มีขั้นตอนย่อยดังต่อไปนี้

**ขั้นตอนที่ 1.1** สำหรับทุกๆ attribute $A_p$ ที่ $p = 1$ ถึง $x$

**ขั้นตอนที่ 1.2 ถึง 1.3** กำหนดให้ $d_1$ เป็นค่าน้อยที่สุด และ $d_n$ เป็นค่ามากที่สุดของ attribute $A_p$ จากตัวอย่างของภาพประกอบ 3.3 ใน step 1.1 จะได้ $d_1$ มีค่าเท่ากับ 1 และ $d_n$ มีค่าเท่ากับ 35

**ขั้นตอนที่ 1.4** เรียงลำดับของข้อมูล attribute $A_p$ จากน้อยไปหามาก แสดงดังตัวอย่างใน step 1.3

**ขั้นตอนที่ 1.5** ทำการ Merge หรือจัดกลุ่มข้อมูลที่มีคลาสเดียวกันที่อยู่ติดกัน ให้อยู่ใน ช่วงข้อมูลเดียวกัน แสดงดังตัวอย่างใน step 1.5 เช่น interval 1 = [1-3], interval 2 = [4 – 7] เป็นต้น

**ขั้นตอนที่ 1.6 ถึง 1.7** วนซ้ำที่ $r = 1$ ถึง $n$ เพื่อคำนวณหาค่ากลาง (midpoint) ของแต่ละช่วงข้อมูลเพื่อกำหนดเป็นขอบเขตข้อมูลในแต่ละช่วงข้อมูล ซึ่งสามารถคำนวณได้โดยใช้สมการที่ (3.1) กำหนดให้ $B_{+r}$ คือขอบเขตของแต่ละช่วงข้อมูล $Bmax_r$ คือขอบบนของช่วงข้อมูลที่ r และ $Bmin_{r+1}$ คือขอบล่างของช่วงข้อมูลที่ r+1 แล้วหารด้วยสองเพื่อหาค่ากลางระหว่างสองช่วงข้อมูลที่อยู่ติดกัน แสดงดังตัวอย่างใน step 1.4 เช่น ค่า midpoint ของช่วงข้อมูลที่ 1 กับช่วงข้อมูลที่ 2 คือ (3+4 ) / 2 = 3.5 เป็นต้น

$$B_{+r} = (Bmax_r + Bmin_{r+1})/2 \tag{3.1}$$

**ขั้นตอนที่ 1.8** แสดงผลลัพธ์ที่ได้จากขั้นตอนที่ 1 จะได้ข้อมูลที่จัดอยู่ในรูปแบบของตาราง 2D quanta matrix ดังแสดงในตารางที่ 3.4 ประกอบด้วย 30 instances, 9 ช่วงข้อมูล และ 3 คลาส

ตารางที่ 3.4 ตัวอย่างผลลัพธ์ 2D quanta matrix ของชุดข้อมูลอายุที่ได้จากขั้นตอนที่ 1

| Class | Interval | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| | [1-3.5] | (3.5-8] | (8-9.5] | (9.5-11.5] | (11.5-13.5] | (13.5-23.5] | (23.5-29.5] | (29.5-32] | (32-35] | |
| Care | 8 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| Edu | 0 | 5 | 0 | 3 | 0 | 5 | 0 | 1 | 0 | 14 |
| Work | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 2 | 7 |
| Total | 8 | 5 | 1 | 3 | 1 | 5 | 4 | 1 | 2 | 30 |

**ขั้นตอนที่ 2** เป็นขั้นตอนในการ Merge โดยใช้ Class Attribute Interval Average (CAIA) ตัวอย่างการทำงานของขั้นตอนวิธีนี้แสดงดังตารางที่ 3.5-3.11 โดยมีรายละเอียดดังต่อไปนี้

**ขั้นตอนที่ 2.1 ถึง 2.3** ทำซ้ำในแต่ละ Class (i = 1 to s) เพื่อคำนวณหาค่า $CAI_{+r}$ ของทุกๆ ช่วงข้อมูล (r =1 to n) โดยใช้สมการที่ (3.2) กำหนดให้ $CAI_{+r}$ เป็นค่าการกระจายตัวของข้อมูลในทุกๆ ช่วงข้อมูลที่ r ของทุกๆ class โดยใช้ $q_{ir}{}^2$ หารด้วย $M_{i+}$ ที่คูณด้วยค่าน้ำหนัก (weight) ที่เป็นผลรวมของแต่ละช่วงข้อมูล ที่ r ในทุกๆ คลาส ถ้าค่าการกระจายตัวของข้อมูลในช่วงข้อมูลที่ r มีค่าต่ำ (ค่า $CAI_{+r}$ ต่ำ) แสดงว่า ช่วงข้อมูลนั้นมีความเป็นอิสระจากคลาสมาก หมายถึงสามารถรวมช่วงข้อมูลนี้กับช่วงอื่นๆ ได้โดยไม่มีผลมากนัก แต่ถ้าค่าการกระจายตัวของข้อมูลในช่วงข้อมูลที่ r นั้นสูง (ค่า $CAI_{+r}$ สูง) แสดงว่าช่วงข้อมูลนั้นมีความเป็นอิสระจากคลาสน้อย หรืออีกนัยหนึ่งคือ ช่วงข้อมูลนี้เหมาะสมสำหรับคลาสนั้นๆ แล้ว ตัวอย่างการคำนวณของ $CAI_{+r}$ ในแต่ละช่วงข้อมูลที่ r แสดงดังตารางที่ 3.5 โดยค่า $CAI_{+r}$ ของช่วงที่ 1 ได้ค่าสูงสุดคือ 56.889 หมายถึงช่วงข้อมูลนี้ไม่จำเป็นต้องมีการรวมกับช่วงข้อมูลอื่นๆ เนื่องจากช่วงข้อมูลนี้ขึ้นอยู่กับคลาสที่เหมาะสมแล้ว ส่วนช่วงข้อมูลที่ 8 $CAI_{+r}$ มีค่าน้อยที่สุดคือ 0.071 หมายถึง ช่วงข้อมูลนี้มีความเป็นอิสระจากคลาสสูงสุด และเหมาะสมที่จะเป็นช่วงข้อมูลที่ถูกหลอมรวมกันมากที่สุด เพราะช่วงข้อมูลนี้มีความเป็นอิสระจากคลาสสูง ทำให้มีผลกระทบต่อช่วงข้อมูลที่รวมกันน้อยที่สุด

$$CAI_{+r} = \sum_{i=1}^{s} (q_{ir}{}^2 / M_{i+}) \times (M_{+r}) \tag{3.2}$$

$$CAIA = \sum_{r=1}^{n} CAI_{+r} / n \tag{3.3}$$

ตารางที่ 3.5 ตัวอย่างการคำนวณหาค่า $CAI_{+r}$ และ CAIA ของชุดข้อมูลอายุในรอบที่ 1

| Class | Interval | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | |
| | [1-3.5] | (3.5-8] | (8-9.5] | (9.5-11.5] | (11.5-13.5] | (13.5-23.5] | (23.5-29.5] | (29.5-32] | (32-35] | |
| Care | 8 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| Edu | 0 | 5 | 0 | 3 | 0 | 5 | 0 | 1 | 0 | 14 |
| Work | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 2 | 7 |
| Total | 8 | 5 | 1 | 3 | 1 | 5 | 4 | 1 | 2 | 30 |
| $CAI_{+r}$ | 56.8 | 8.92 | 0.11 | 1.92 | 0.14 | 8.92 | 9.14 | **0.07** | 1.14 | CAIA=9.698 |

**ขั้นตอนที่ 2.4** ตรวจเงื่อนไขที่ใช้ในพิจารณาการหยุดคือ ถ้าจำนวนช่วงข้อมูล k ที่ได้ มากกว่าจำนวนคลาส s ก็จะทำการหลอมรวมต่อไป จนกระทั่งจำนวนช่วงข้อมูล k ที่ได้จะน้อยกว่าหรือเท่ากับจำนวนคลาส s ให้หยุดการหลอมรวม

**ขั้นตอนที่ 2.5** ค้นหาช่วงข้อมูลที่มีค่าของ $CAI_{+r}$ ที่น้อยที่สุด และกำหนดให้ค่า $CAI_{min}$ เป็นช่วงข้อมูลที่มีค่า $CAI_{+r}$ ที่น้อยที่สุด

**ขั้นตอนที่ 2.6 ถึง 2.7** ทดสอบหลอมรวมกับช่วงข้อมูลที่อยู่ติดกันทางขวาของช่วงข้อมูลที่ $CAI_{min}$ แล้วทำการคำนวณหาค่า $CAIA_{right}$ โดยใช้สมการที่ (3.3) กำหนดให้ CAIA เป็นค่าเฉลี่ยของการกระจายตัวในทุกๆ ช่วงข้อมูลที่มีการหลอมรวม ถ้าค่าเฉลี่ยของ CAIA สูงกว่า ก็แสดงว่าค่าเฉลี่ยจากการหลอมรวมกันมีการกระจายตัวที่ดีกว่า ตัวอย่างในตารางที่ 3.5 นั้นคือ ช่วงข้อมูลที่ 8 ได้ค่า $CAI_{+r}$ น้อยสุดคือ 0.071 กำหนดให้เป็น $CAI_{min}$ และ ช่วงข้อมูลที่ 9 เป็นช่วงข้อมูลที่อยู่ติดกันทางขวา หลังจากที่มีการหลอมรวมกันระหว่าง ช่วงข้อมูลที่ 8 และ 9 ค่า $CAIA_{right}$ ที่ได้คือ 11.00

**ขั้นตอนที่ 2.8 ถึง 2.9** ทดสอบหลอมรวมกับช่วงข้อมูลที่อยู่ติดกันทางซ้ายของ $CAI_{min}$ แล้วทำการคำนวณหาค่า $CAIA_{left}$ ตัวอย่างในตารางที่ 3.5 ช่วงข้อมูลที่อยู่ทางซ้ายของ $CAI_{min}$ คือ ช่วงข้อมูลที่ 7 หลังจากที่มีการหลอมรวมกันระหว่าง ช่วงข้อมูลที่ 7 และ 8 ค่า $CAIA_{left}$ ที่ได้คือ 11.232

**ขั้นตอนที่ 2.10 ถึง 2.12** เปรียบเทียบค่า CAIA ของการหลอมรวมกับทางขวา (ช่วงข้อมูล 8 กับ 9 ได้ค่า CAIA คือ 11.00) และทางซ้าย (ช่วงข้อมูล 7 กับ 8 ได้ค่า CAIA คือ 11.232)

แล้วเลือกหลอมรวมกับค่า CAIA ที่มากกว่า ดังนั้นค่าที่หลอมรวมทางซ้ายจะได้ค่า CAIA มากกว่าทางขวา จึงทำการเลือกหลอมรวมกับช่วงข้อมูลทางซ้ายนั้นคือ (ช่วงข้อมูลที่ 7 กับ 8) แล้วอัพเดทค่า CAIA เป็น 11.232

**ขั้นตอนที่ 2.13** set ค่า n = n-1 แล้วกลับไปทำซ้ำขั้นตอนที่ 2.4 แสดงดังตารางที่ 3.6 ถึง 3.10

**ขั้นตอนที่ 2.14** แสดงผลลัพธ์ของช่วงข้อมูลที่ได้มีการแบ่งช่วงข้อมูลดังตัวอย่างในตารางที่ 3.11

ตารางที่ 3.6 ตัวอย่างการคำนวณหาค่า $CAI_{+r}$ และ CAIA ของชุดข้อมูลอายุในรอบที่ 2

| Class | Interval | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | |
| | [1-3.5] | (3.5-8] | (8-9.5] | (9.5-11.5] | (11.5-13.5] | (13.5-23.5] | (23.5-32] | (32-35] | |
| Care | 8 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 9 |
| Edu | 0 | 5 | 0 | 3 | 0 | 5 | 1 | 0 | 14 |
| Work | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 2 | 7 |
| Total | 8 | 5 | 1 | 3 | 1 | 5 | 5 | 2 | 30 |
| **$CAI_{+r}$** | 56.88 | 8.92 | **0.11** | 1.92 | 0.14 | 8.92 | 11.78 | 1.14 | CAIA=11.232 |

จากตารางที่ 3.6 เป็นตัวอย่างการทำงานของ CAIA ในวนรอบที่ 2 คือจะทำการค้นหาช่วงข้อมูลที่มีค่า $CAI_{+r}$ ที่น้อยที่สุด คือช่วงข้อมูลที่ 3 โดยมีค่า $CAI_{+r}$ คือ 0.11 แล้วทดสอบการหลอมรวมกับช่วงข้อมูลทางซ้ายคือ ช่วงข้อมูลที่ 2 จะได้ค่า CAIA คือ 13.171 และทดสอบการหลอมรวมกับช่วงข้อมูลทางขวา คือ ช่วงข้อมูลที่ 4 จะได้ค่า CAIA คือ 12.976 แล้วเปรียบเทียบค่า CAIA ของทั้งสองช่วงข้อมูล โดยทำการเลือกค่า CAIA ที่สูงกว่า จากตัวอย่างนี้จะต้องเลือกหลอมรวมกันของช่วงข้อมูลที่ 2 และ 3 นั้นก็คือ (3.5-8] กับ (8-9.5] จะได้เป็น (3.5-9.5] ไปเป็นช่วงข้อมูลที่ 2 ในตารางที่ 3.7

ตารางที่ 3.7 ตัวอย่างการคำนวณหาค่า $CAI_{+r}$ และ CAIA ของชุดข้อมูลอายุในรอบที่ 3

| Class | Interval | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| | [1-3.5] | (3.5-9.5] | (9.5-11.5] | (11.5-13.5] | (13.5-23.5] | (23.5-32] | (32-35] | |
| Care | 8 | 1 | 0 | 0 | 0 | 0 | 0 | 9 |
| Edu | 0 | 5 | 3 | 0 | 5 | 1 | 0 | 14 |
| Work | 0 | 0 | 0 | 1 | 0 | 4 | 2 | 7 |
| Total | 8 | 6 | 3 | 1 | 5 | 5 | 2 | 30 |
| **$CAI_{+r}$** | 56.88 | 11.38 | 1.92 | **0.14** | 8.92 | 11.78 | 1.14 | CAIA=13.171 |

จากตารางที่ 3.7 เป็นตัวอย่างการทำงานของ CAIA ในวนรอบที่ 3 คือจะทำการค้นหาช่วงข้อมูลที่มีค่า $CAI_{+r}$ ที่น้อยที่สุด คือช่วงข้อมูลที่ 4 โดยมีค่า $CAI_{+r}$ คือ 0.14 แล้วทดสอบการหลอมรวมกับช่วงข้อมูลทางซ้ายคือ ช่วงข้อมูลที่ 3 จะได้ค่า CAIA คือ 15.544 และทดสอบการหลอมรวมกับช่วงข้อมูลทางขวา คือ ช่วงข้อมูลที่ 5 จะได้ค่า CAIA คือ 15.783 แล้วเปรียบเทียบค่า CAIA ของทั้งสองช่วงข้อมูล โดยทำการเลือกค่า CAIA ที่สูงกว่า จากตัวอย่างนี้จะต้องเลือกหลอมรวมกันของช่วงข้อมูลที่ 4 และ 5 นั้นก็คือ (11.5-13.5] กับ (13.5-23.5] จะได้เป็น (11.5-23.5] ไปเป็นช่วงข้อมูลที่ 4 ในตารางที่ 3.8

ตารางที่ 3.8 ตัวอย่างการคำนวณหาค่า $CAI_{+r}$ และ CAIA ของชุดข้อมูลอายุในรอบที่ 4

| Class | Interval | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | |
| | [1-3.5] | (3.5-9.5] | (9.5-11.5] | (11.5-23.5] | (23.5-32] | (32-35] | |
| Care | 8 | 1 | 0 | 0 | 0 | 0 | 9 |
| Edu | 0 | 5 | 3 | 5 | 1 | 0 | 14 |
| Work | 0 | 0 | 0 | 1 | 4 | 2 | 7 |
| Total | 8 | 6 | 3 | 6 | 5 | 2 | 30 |
| **$CAI_{+r}$** | 56.88 | 11.38 | 1.92 | 11.57 | 11.78 | **1.14** | CAIA=15.783 |

จากตารางที่ 3.8 เป็นตัวอย่างการทำงานของ CAIA ในวนรอบที่ 4 คือจะทำการค้นหาช่วงข้อมูลที่มีค่า $CAI_{+r}$ ที่น้อยที่สุด คือช่วงข้อมูลที่ 6 โดยมีค่า $CAI_{+r}$ คือ 1.14 กรณีนี้ช่วงข้อมูลที่ 6 อยู่ช่วงข้อมูลสุดท้าย ดังนั้นจึงต้องหลอมรวมกับช่วงข้อมูลที่ 5 นั้นก็คือ (23.5-32] กับ (32-35] จะได้เป็น (23.5-35] ไปเป็นช่วงข้อมูลที่ 5 ในตารางที่ 3.9

ตารางที่ 3.9 ตัวอย่างการคำนวณหาค่า $CAI_{+r}$ และ CAIA ของชุดข้อมูลอายุในรอบที่ 5

| Class | Interval | | | | | Total |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| | [1-3.5] | (3.5-9.5] | (9.5-11.5] | (11.5-23.5] | (23.5-35] | |
| Care | 8 | 1 | 0 | 0 | 0 | 9 |
| Edu | 0 | 5 | 3 | 5 | 1 | 14 |
| Work | 0 | 0 | 0 | 1 | 6 | 7 |
| Total | 8 | 6 | 3 | 6 | 7 | 30 |
| **$CAI_{+r}$** | 56.88 | 11.38 | **1.92** | 11.57 | 36.50 | CAIA=23.653 |

จากตารางที่ 3.9 เป็นตัวอย่างการทำงานของ CAIA ในวนรอบที่ 5 ช่วงข้อมูลที่มีค่า $CAI_{+r}$ ที่น้อยที่สุด คือช่วงข้อมูลที่ 3 โดยมีค่า $CAI_{+r}$ คือ 1.92 แล้วทดสอบการหลอมรวมกับช่วงข้อมูลทางซ้ายคือ ช่วงข้อมูลที่ 2 จะได้ค่า CAIA คือ 36.775 และทดสอบการหลอมรวมกับช่วงข้อมูลทางขวา คือ ช่วงข้อมูลที่ 4 จะได้ค่า CAIA คือ 36.799 แล้วเปรียบเทียบค่า CAIA ของทั้งสองช่วงข้อมูล โดยทำการเลือกค่า CAIA ที่สูงกว่า จากตัวอย่างนี้จะต้องเลือกหลอมรวมกันของช่วงข้อมูลที่ 3 และ 4 นั้นก็คือ (9.5-11.5] กับ (11.5-23.5] จะได้เป็น (9.5-23.5] ไปเป็นช่วงข้อมูลที่ 3 ในตารางที่ 3.10

ตารางที่ 3.10 ตัวอย่างการคำนวณหาค่า $CAI_{+r}$ และ CAIA ของชุดข้อมูลอายุในรอบที่ 6

| Class | Interval | | | | Total |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |
| | [1-3.5] | (3.5-9.5] | (9.5-23.5] | (23.5-32] | |
| Care | 8 | 1 | 0 | 0 | 9 |
| Edu | 0 | 5 | 8 | 1 | 14 |
| Work | 0 | 0 | 1 | 6 | 7 |
| Total | 8 | 6 | 9 | 7 | 30 |
| **$CAI_{+r}$** | 56.88 | **11.38** | 42.42 | 36.50 | CAIA=36.799 |

จากตารางที่ 3.10 เป็นตัวอย่างการทำงานของ CAIA ในวนรอบที่ 6 ช่วงข้อมูลที่มีค่า $CAI_{+r}$ ที่น้อยที่สุด คือช่วงข้อมูลที่ 2 โดยมีค่า $CAI_{+r}$ คือ 11.38 แล้วทดสอบการหลอมรวมกับช่วงข้อมูลทางซ้ายคือ ช่วงข้อมูลที่ 1 จะได้ค่า CAIA คือ 76.642 และทดสอบการหลอมรวมกับช่วงข้อมูลทางขวา คือ ช่วงข้อมูลที่ 3 จะได้ค่า CAIA คือ 92.756 แล้วเปรียบเทียบค่า CAIA ของทั้งสองช่วงข้อมูล โดยทำการเลือกค่า CAIA ที่สูงกว่า จากตัวอย่างนี้จะต้องเลือกหลอมรวมกันของช่วงข้อมูลที่ 2 และ 3 นั้นก็คือ (3.5-9.5] กับ (9.5-23.5] จะได้เป็น (3.5-23.5] เมื่อเข้าวนรอบที่ 7 จำนวนของช่วงข้อมูลที่ได้คือ 3 ซึ่งเท่ากับจำนวนของคลาส ดังนั้นขั้นตอนวิธีของ CAIA จึงหยุดทำการแบ่งช่วงข้อมูลโดยจะได้ผลลัพธ์ดังแสดงในตารางที่ 3.11

ตารางที่ 3.11 ตัวอย่างผลลัพธ์ของการแบ่งช่วงข้อมูลได้จากชุดข้อมูลอายุโดยใช้ขั้นตอนวิธี CAIA

| Class | Interval | | | Total |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| | [1-3.5] | (3.5-23.5] | (23.5-32] | |
| Care | 8 | 1 | 0 | 9 |
| Edu | 0 | 13 | 1 | 14 |
| Work | 0 | 1 | 6 | 7 |
| Total | 8 | 15 | 7 | 30 |
| **$CAI_{+r}$** | 56.88 | 184.88 | 36.50 | CAIA=92.756 |

สำหรับจำนวนวนรอบและช่วงข้อมูลที่มีการหลอมรวมกันจะหยุดเมื่อ n น้อยกว่าหรือเท่ากับจำนวนของคลาส ดังนั้นตัวอย่างชุดข้อมูลอายุโดยใช้ขั้นตอนวิธีในการแบ่งช่วงข้อมูลของ CAIA จะหยุดเมื่อ n มีค่าเท่ากับ 3 ซึ่งสามารถสรุปรายละเอียดได้ดังตารางที่ 3.12

ตารางที่ 3.12 สรุปจำนวนของวนรอบที่ใช้ในการแบ่งช่วงข้อมูลโดยใช้ขั้นตอนวิธี CAIA

| **Loop** | **Interval merge** | **CAIA value** |
|---|---|---|
| 1 | (23.5-29.5] merge (29.5-32] | 11.232 |
| 2 | (3.5-8] merge (8-9.5] | 13.171 |
| 3 | (11.5-13.5] merge (13.5-23.5] | 15.783 |
| 4 | (23.5-32] merge (32-35] | 23.653 |
| 5 | (9.5-11.5] merge (11.5-23.5] | 36.799 |
| 6 | (3.5-11.5] merge (11.5-23.5] | 92.757 |

# บทที่ 4

## ผลการทดลองและวิจารณ์

บทนี้จะนำเสนอผลลัพธ์ที่ได้จากการทดลองตามที่ได้ออกแบบและพัฒนาขั้นตอนวิธีที่ใช้ในการแบ่งช่วงข้อมูลโดยใช้ค่าเฉลี่ยการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ (Class Attribute Interval Average: CAIA) เป็นเกณฑ์ในการพิจารณาช่วงข้อมูลที่ดีที่สุดในการหลอมรวมของช่วงข้อมูลที่อยู่ติดกัน ในการทดลองเพื่อวัดประสิทธิภาพในการทำงานของขั้นตอนวิธีนี้ ได้ทดลองกับชุดข้อมูล 4 ชุดข้อมูล Benchmarks จากฐานข้อมูลของ UCI Data Set (Blake and Merz, 1998) คือ Iris, Breast Cancer, Heart disease และ Glass ผลลัพธ์ของการทดลองจะแบ่งออกเป็น 2 ประเด็นคือ ประเด็นแรก เป็นการวัดประสิทธิภาพในเรื่องของค่าความถูกต้องโดยใช้ 4 Data mining algorithm ในโปรแกรม WEKA (Lan and Frank, 2005) โดยแบ่งชุดข้อมูลในการทดสอบแบบ 10 Fold Cross Validation และประเด็นที่สอง เป็นการวัดจำนวนของช่วงข้อมูลที่ได้จากการแบ่งช่วงข้อมูล โดยในการทดลองประสิทธิภาพนั้นจะเปรียบเทียบกับ 6 Discretization Algorithms คือ Equal-Width (EW) (Wong and Chiu, 1987), Equal-Frequency (EF) (Wong and Chiu, 1987), ChiMerge (Kerber, 1992), Information Entropy Maximization (IEM) (Fayyad and Irani, 1993), Class-Attribute Interdependence Maximization (CAIM) (Kurgan and Cios, 2004) และ Class-Attribute Contingency Coefficient (CACC) (Tsai, 2008)

### 4.1 ชุดข้อมูลที่ใช้ในการทดลอง

ชุดข้อมูลที่ใช้ในการทดสอบประสิทธิภาพของขั้นตอนวิธีในการแบ่งช่วงข้อมูลประกอบด้วย 4 ชุดข้อมูลที่เป็นชุดข้อมูลแบบต่อเนื่อง (Continuous Data) คือ 1) ชุดข้อมูล Iris ประกอบด้วย จำนวนแถวข้อมูล (Instance) 150 แถวข้อมูล มีจำนวนแอตทริบิวต์ 4 แอตทริบิวต์ และมีจำนวนคลาส 3 คลาส 2) ชุดข้อมูล Breast Cancer ประกอบด้วย จำนวนแถวข้อมูล 699 แถวข้อมูล มีจำนวนแอตทริบิวต์ 10 แอตทริบิวต์ และมีจำนวนคลาส 2 คลาส 3) ชุดข้อมูล Heart Disease ประกอบด้วย จำนวนแถวข้อมูล 303 แถวข้อมูล มีจำนวนแอตทริบิวต์ 13 แอตทริบิวต์ และมีจำนวนคลาส 5 คลาส และ 4) ชุดข้อมูล Glass ประกอบด้วย จำนวนแถวข้อมูล 214 แถวข้อมูล มีจำนวนแอตทริบิวต์ 9 แอตทริบิวต์ และมีจำนวนคลาส 6 คลาส

### 4.1.1 ชุดข้อมูล Iris

สำหรับคุณลักษณะของชุดข้อมูล Iris แสดงตัวอย่างดังตารางที่ 4.1 ประกอบด้วยจำนวนแถวข้อมูล (Instance) 150 แถวข้อมูล มีจำนวนแอตทริบิวต์ทั้งหมด 4 แอตทริบิวต์ คือ 1) Sepal Length มีหน่วยเป็น (cm) 2) Sepal Width มีหน่วยเป็น (cm) 3) Petal Length มีหน่วยเป็น (cm) และ 4) Petal Width มีหน่วยเป็น (cm) และมีจำนวนคลาส 3 คลาส คือ Iris Setosa, Iris Versicolour และ Iris Virginica โดยในแต่ละคลาสจะมีข้อมูล 50 แถวข้อมูลเท่าๆ กันในทุกๆ คลาส และในตารางที่ 4.2 แสดงตัวอย่างของชุดข้อมูล Iris ที่ใช้ในการทดลอง

ตารางที่ 4.1 คุณลักษณะของชุดข้อมูล Iris ที่ใช้ในการทดลอง

| ชื่อชุดข้อมูล | จำนวนแถวข้อมูล | จำนวนแอตทริบิวต์ | จำนวนคลาส | ค่าข้อมูลสูญหาย |
|---|---|---|---|---|
| Iris | 150 | X1= Sepal Length<br>X2 = Sepal Width<br>X3 = Petal Length<br>X4 = Petal Width | 1) Iris Setosa<br>2) Iris Versicolour<br>3) Iris Virginica | ไม่มี |

ตารางที่ 4.2 ตัวอย่างชุดข้อมูล Iris ที่ใช้ในการทดลอง

| X1 | X2 | X3 | X4 | Class |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | Iris-Setosa |
| 4.9 | 3 | 1.4 | 0.2 | Iris-Setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | Iris-Setosa |
| 7 | 3.2 | 4.7 | 1.4 | Iris-Versicolor |
| 6.4 | 3.2 | 4.5 | 1.5 | Iris-Versicolor |
| 6.9 | 3.1 | 4.9 | 1.5 | Iris-Versicolor |
| 7.6 | 3 | 6.6 | 2.1 | Iris-Virginica |
| 4.9 | 2.5 | 4.5 | 1.7 | Iris-Virginica |
| 7.3 | 2.9 | 6.3 | 1.8 | Iris-Virginica |

### 4.1.2 ชุดข้อมูล Breast Cancer

สำหรับคุณลักษณะของชุดข้อมูล Breast Cancer แสดงตัวอย่างดังตารางที่ 4.3 ประกอบด้วย จำนวนแถวข้อมูล 699 แถวข้อมูล มีจำนวนแอตทริบิวต์ทั้งหมด 10 แอตทริบิวต์ คือ 1) Sample code number 2) Clump Thickness 3) Uniformity of Cell Size 4) Uniformity of Cell Shape 5) Marginal Adhesion 6) Single Epithelial Cell Size 7) Bare Nuclei 8) Bland Chromatin 9) Normal Nucleoli และ 10) Mitoses และมีจำนวนคลาส 2 คลาส คือ Benign และ Malignant และในตารางที่ 4.4 แสดงตัวอย่างของชุดข้อมูล Breast Cancer ที่ใช้ในการทดลอง

ตารางที่ 4.3 คุณลักษณะของชุดข้อมูล Breast Cancer ที่ใช้ในการทดลอง

| ชื่อชุดข้อมูล | จำนวนแถวข้อมูล | จำนวนแอตทริบิวต์ | จำนวนคลาส | ค่าข้อมูลสูญหาย |
|---|---|---|---|---|
| Breast Cancer | 699 | X1 = Sample code number<br>X2 = Clump Thickness<br>X3 = Uniformity of Cell Size<br>X4 = Uniformity of Cell Shape<br>X5 = Marginal Adhesion<br>X6 =Single Epithelial Cell Size<br>X7 = Bare Nuclei<br>X8 = Bland Chromatin<br>X9 = Normal Nucleoli<br>X10 = Mitoses | 1) Benign<br>2) Malignant | มี |

ตารางที่ 4.4 ตัวอย่างชุดข้อมูล Breast Cancer ที่ใช้ในการทดลอง

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | Benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | Benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | - | Benign |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | Benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | Benign |
| 1054590 | 7 | 3 | 2 | 10 | 5 | 10 | 5 | 4 | 4 | Malignant |
| 1054593 | 10 | 5 | 5 | 3 | 6 | 7 | 7 | 10 | 1 | Malignant |
| 1057013 | 8 | 4 | 5 | 1 | 2 | - | 7 | 3 | 1 | Malignant |
| 1065726 | 5 | 2 | 3 | 4 | 2 | 7 | 3 | 6 | 1 | Malignant |
| 1072179 | 10 | 7 | 7 | 3 | 8 | 5 | 7 | 4 | 3 | Malignant |

### 4.1.3 ชุดข้อมูล Heart Disease

สำหรับคุณลักษณะของชุดข้อมูล Heart Disease แสดงตัวอย่างดังตารางที่ 4.5 ประกอบด้วย จำนวนแถวข้อมูล 303 แถวข้อมูล มีจำนวนแอตทริบิวต์ทั้งหมด 13 แอตทริบิวต์ คือ 1) age 2) sex 3) cp (chest pain type) 4) trestbps (resting blood pressure) 5) chol (serum cholestoral) 6) fbs (fasting blood sugar) 7) restecg (resting electrocardiographic results) 8) thalach (maximum heart rate achieved) 9) exang (exercise induced angina) 10) oldpeak (ST depression induced by exercise relative to rest) 11) slope (the slope of the peak exercise ST segment) 12) ca (number of major vessels (0-3) colored by flourosopy) และ 13) thal (3 = normal; 6 = fixed defect; 7 = reversable defect ) และมีจำนวนคลาส 5 คลาส คือ value = 0 , value = 1, value = 2, value = 3 และ value = 4 และในตารางที่ 4.6 แสดงตัวอย่างของชุดข้อมูล Heart Disease ที่ใช้ในการทดลอง

ตารางที่ 4.5 คุณลักษณะของชุดข้อมูล Heart Disease ที่ใช้ในการทดลอง

| ชื่อชุดข้อมูล | จำนวนแถวข้อมูล | จำนวนแอตทริบิวต์ | จำนวนคลาส | ค่าข้อมูลสูญหาย |
|---|---|---|---|---|
| Heart Disease | 303 | X1 = age<br>X2 = sex<br>X3 = cp<br>X4 = trestbps<br>X5 = chol<br>X6 = fbs<br>X7 = restecg<br>X8 = thalach<br>X9 = exang<br>X10 = oldpeak<br>X11 = slope<br>X12 = ca<br>X13 = thal | value = 0<br>value = 1<br>value = 2<br>value = 3<br>value = 4 | มี |

ตารางที่ 4.6 ตัวอย่างชุดข้อมูล Heart Disease ที่ใช้ในการทดลอง

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 3 | 0 | 6 | 0 |
| 37 | 1 | 3 | 130 | 250 | 0 | 0 | 187 | 0 | 3.5 | 3 | 0 | 3 | 0 |
| 41 | 0 | 2 | 130 | 204 | 0 | 2 | 172 | 0 | 1.4 | 1 | 0 | 3 | 0 |
| 56 | 1 | 2 | 120 | 236 | 0 | 0 | 178 | 0 | 0.8 | 1 | 0 | - | 0 |
| 57 | 0 | 4 | 120 | 354 | 0 | 0 | 163 | 1 | 0.6 | 1 | 0 | 3 | 0 |
| 61 | 0 | 4 | 130 | 330 | 0 | 2 | 169 | 0 | 0 | 1 | 0 | 3 | 1 |
| 60 | 1 | - | 130 | 253 | 0 | 0 | 144 | 1 | 1.4 | 1 | 1 | 7 | 1 |
| 54 | 1 | 4 | 124 | 266 | 0 | 2 | 109 | 1 | 2.2 | 2 | 1 | 7 | 1 |

### 4.1.4 ชุดข้อมูล Glass

สำหรับคุณลักษณะของชุดข้อมูล Glass แสดงตัวอย่างดังตารางที่ 4.7 ประกอบด้วยจำนวนแถวข้อมูล 214 แถวข้อมูล มีจำนวนแอตทริบิวต์ทั้งหมด 9 แอตทริบิวต์ คือ 1) RI (refractive index) 2) Na (Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10) 3) Mg (Magnesium) 4) Al (Aluminum) 5) Si (Silicon) 6) K (Potassium) 7) Ca (Calcium) 8) Ba (Barium) และ 9) Fe (Iron) และมีจำนวนคลาส 6 คลาส คือ 1) building_windows_float 2) building_windows_non 3) vehicle_windows_float 4) containers 5) Tableware และ 6) Headlamps และในตารางที่ 4.8 แสดงตัวอย่างของชุดข้อมูล Glass ที่ใช้ในการทดลอง

ตารางที่ 4.7 คุณลักษณะของชุดข้อมูล Glass ที่ใช้ในการทดลอง

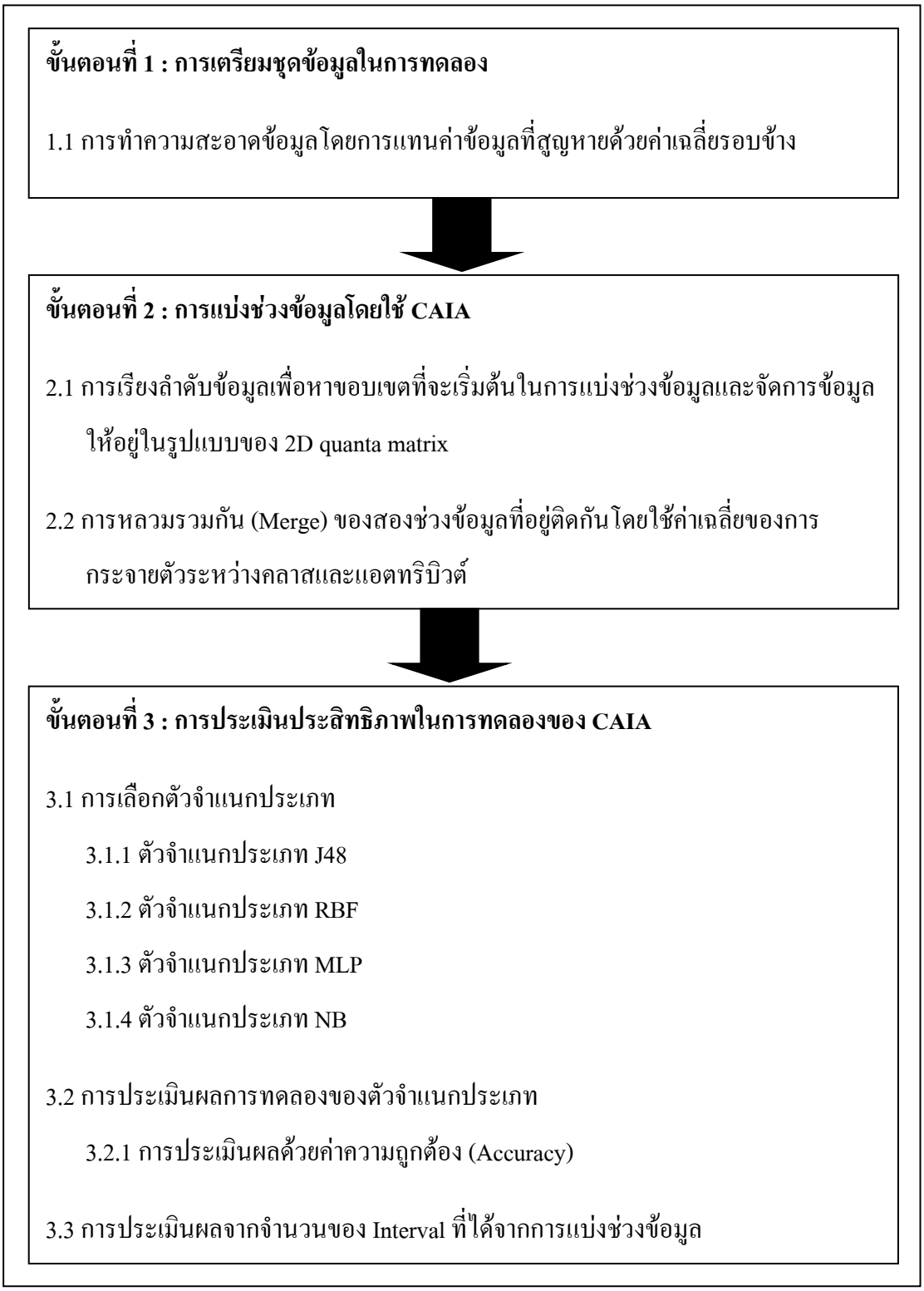
| ชื่อชุดข้อมูล | จำนวนแถวข้อมูล | จำนวนแอตทริบิวต์ | จำนวนคลาส | ค่าข้อมูลสูญหาย |
|---|---|---|---|---|
| Glass | 214 | X1 = RI<br>X2 = Na<br>X3 = Mg<br>X4 = Al<br>X5 = Si<br>X6 = K<br>X7 = Ca<br>X8 = Ba<br>X9 = Fe | 1) building_windows_float<br>2) building_windows_non<br>3) vehicle_windows_float<br>4) containers<br>5) tableware<br>6) headlamps | ไม่มี |

ตารางที่ 4.8 ตัวอย่างชุดข้อมูล Glass ที่ใช้ในการทดลอง

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.51793 | 12.79 | 3.5 | 1.12 | 73.03 | 0.64 | 8.77 | 0 | 0 | 'build wind float' |
| 2 | 1.51643 | 12.16 | 3.52 | 1.35 | 72.89 | 0.57 | 8.53 | 0 | 0 | 'vehic wind float' |
| 3 | 1.51793 | 13.21 | 3.48 | 1.41 | 72.64 | 0.59 | 8.43 | 0 | 0 | 'build wind float' |
| 4 | 1.51299 | 14.4 | 1.74 | 1.54 | 74.55 | 0 | 7.59 | 0 | 0 | tableware |
| 5 | 1.53393 | 12.3 | 0 | 1 | 70.16 | 0.12 | 16.19 | 0 | 0.24 | 'build wind non-float' |
| 6 | 1.51655 | 12.75 | 2.85 | 1.44 | 73.27 | 0.57 | 8.79 | 0.11 | 0.22 | 'build wind non-float' |
| 7 | 1.51779 | 13.64 | 3.65 | 0.65 | 73 | 0.06 | 8.93 | 0 | 0 | 'vehic wind float' |
| 8 | 1.51837 | 13.14 | 2.84 | 1.28 | 72.85 | 0.55 | 9.07 | 0 | 0 | 'build wind float' |
| 9 | 1.51545 | 14.14 | 0 | 2.68 | 73.39 | 0.08 | 9.07 | 0.61 | 0.05 | headlamps |
| 10 | 1.51789 | 13.19 | 3.9 | 1.3 | 72.33 | 0.55 | 8.44 | 0 | 0.28 | 'build wind non-float' |

## 4.2 การทดลองการแบ่งช่วงข้อมูลโดยใช้วิธีการของ CAIA

การทดลองการแบ่งช่วงข้อมูลโดยใช้ค่าเฉลี่ยของกระจายตัวระหว่างคลาสและแอตทริบิวต์ หรือ CAIA มีขั้นตอนวิธีในการทดลองคือ 1) ขั้นตอนการเตรียมข้อมูล 2) ขั้นตอนการแบ่งช่วงข้อมูล และ 3) ขั้นตอนการประเมินประสิทธิภาพในการทดลองของ CAIA โดยมีรายละเอียดในการทดลองดังภาพประกอบ 4.1




ภาพประกอบ 4.1 ขั้นตอนในการทดลอง

**4.2.1 การทดลองชุดข้อมูล Iris**

**ขั้นตอนที่ 1 การเตรียมข้อมูลในการทดลอง**

1.1 นำชุดข้อมูล Iris ที่ดาวน์โหลดมาจาก UCI database มาทำความสะอาดข้อมูล สำหรับชุดข้อมูล Iris เป็นชุดข้อมูลที่มีจำนวนแถวข้อมูล 150 แถวข้อมูล มีจำนวน 4 แอตทริบิวต์ มีคลาสจำนวน 3 คลาส และไม่มีข้อมูลที่สูญหาย ดังนั้นจึงไม่จำเป็นต้องมีการแทนค่าข้อมูลที่สูญหาย

**ขั้นตอนที่ 2 การแบ่งช่วงข้อมูลโดยใช้ CAIA**

2.1 นำชุดข้อมูล Iris ที่ได้จากขั้นตอนการเตรียมข้อมูลมาหาค่าน้อยที่สุดกำหนดเป็น $(d_1)$ และค่าน้อยที่สุดกำหนดเป็น $(d_n)$ แล้วเรียงลำดับข้อมูลจากน้อยไปหามาก

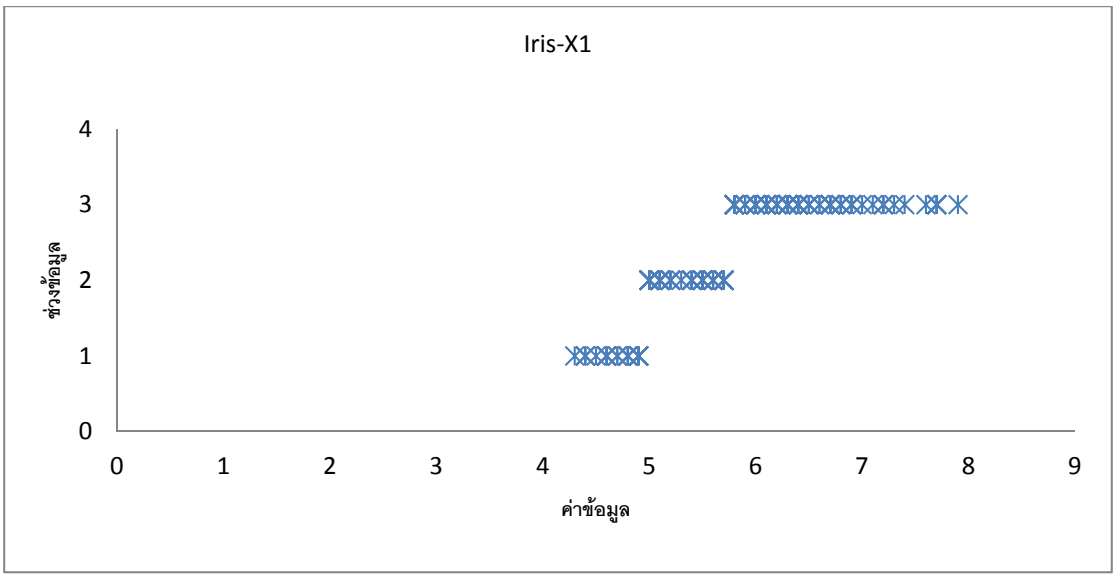
2.2 จัดกลุ่มข้อมูลของค่าข้อมูลที่มีคลาสเดียวกันที่อยู่ติดกัน แล้วหาค่ากลางของแต่ละกลุ่มข้อมูลที่อยู่ติดกันโดยใช้สมการที่ (3.1)

2.3 เซตข้อมูลให้อยู่รูปของตาราง 2D Quanta matrix เพื่อใช้หาค่าความสัมพันธ์ที่กระจายตัวในแต่ละช่วงข้อมูลระหว่างคลาสและแอตทริบิวต์

2.4 หาค่าการกระจายตัวของข้อมูลในแต่ละช่วงข้อมูลโดยใช้สมการที่ (3.2) แล้วหาช่วงข้อมูลที่มีค่าการกระจายตัวน้อยที่สุดเพื่อใช้ในการหลอมรวมกันกับช่วงข้อมูลที่อยู่ติดกัน

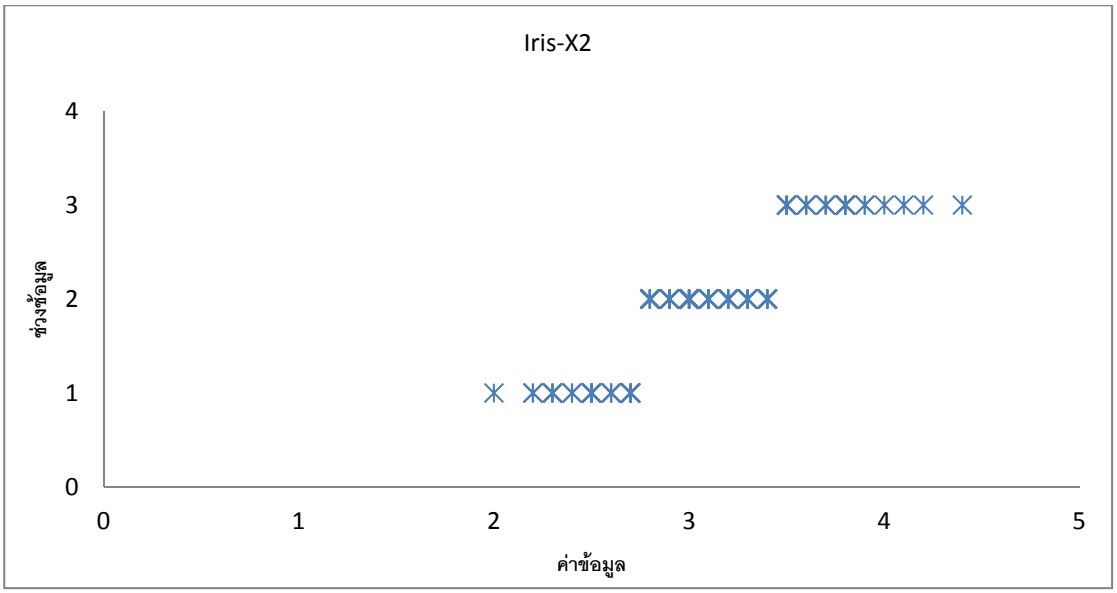
2.5 ทดสอบการหลอมรวมกันกับช่วงข้อมูลที่อยู่ทางซ้ายและทางขวา แล้วหาค่าเฉลี่ยของค่าการกระจายตัวทั้งช่วงข้อมูลอยู่ทางซ้ายและช่วงข้อมูลที่อยู่ทางขวาด้วยสมการที่ (3.3) และเปรียบเทียบค่าเฉลี่ยที่ได้ ถ้าค่าเฉลี่ยของช่วงข้อมูลใดที่มีการหลอมรวมกันแล้วมีค่ามากกว่า ก็ให้เลือกหลอมรวมกับช่วงข้อมูลนั้น จึงกว่าจำนวนของช่วงข้อมูลที่ได้จะเท่ากับจำนวนของคลาสนั้นก็คือ 3

ผลของการแบ่งช่วงข้อมูลชุดข้อมูล Iris สามารถแบ่งตามแอตทริบิวต์ได้ 4 แอตทริบิวต์คือ X1= sepal length X2 = sepal width X3 = petal length และ X4 = petal width มีรายละเอียดดังภาพประกอบ 4.2-4.5

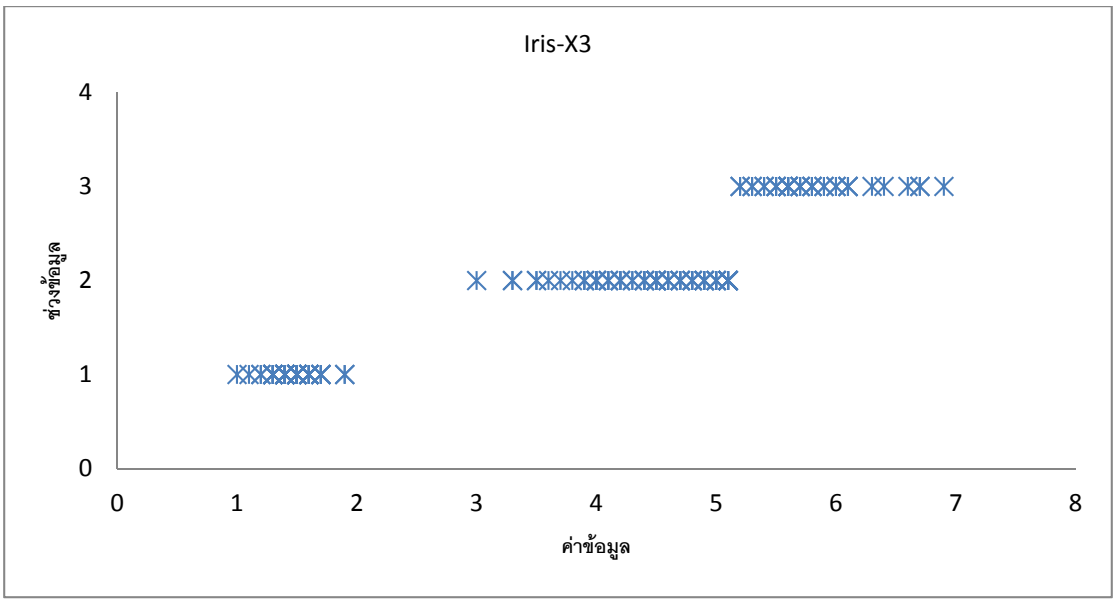



ภาพประกอบ 4.2 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X1 โดยใช้ CAIA

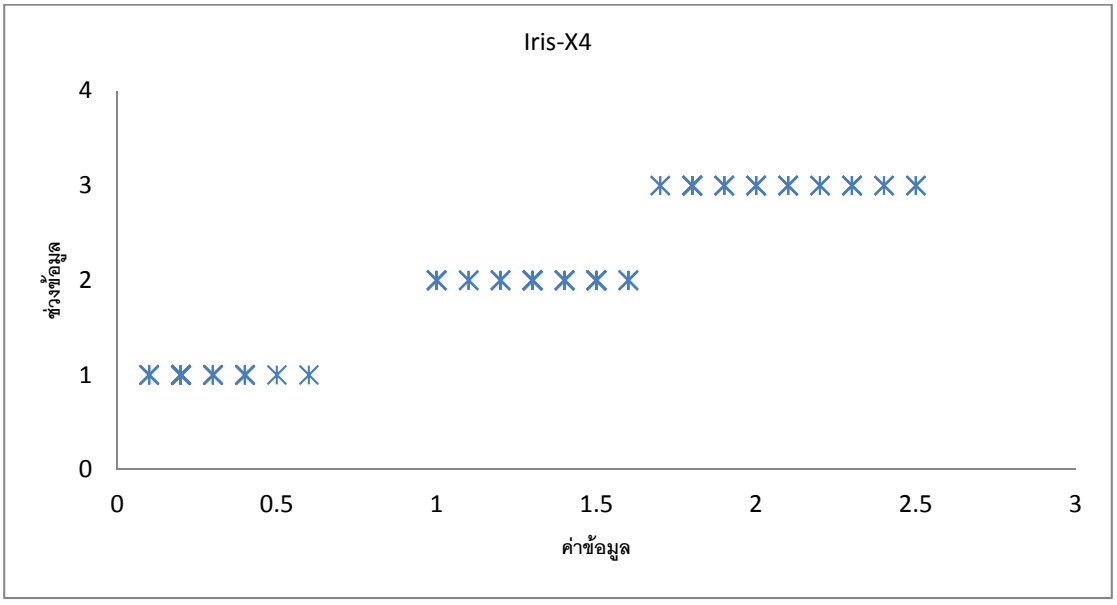
จากภาพประกอบ 4.2 เป็นการแสดงผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลโดยใช้วิธีการของ CAIA แนวตั้งแสดงช่วงข้อมูลที่ได้ และแนวนอนเป็นค่าข้อมูลดิบ ดังนั้นจำนวนช่วงข้อมูลที่ได้คือ 3 ช่วงข้อมูล




ภาพประกอบ 4.3 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X2 โดยใช้ CAIA




ภาพประกอบ 4.4 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X3 โดยใช้ CAIA




ภาพประกอบ 4.5 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X4 โดยใช้ CAIA

**ขั้นตอนที่ 3 การประเมินประสิทธิภาพในการทดลองของ CAIA**

นำชุดข้อมูล Iris ที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ไปทดสอบประสิทธิภาพกับตัวจำแนกประเภทในโปรแกรม WEKA คือ J48 RBF MLP และ NB โดยใช้ค่าความถูกต้องเป็นเกณฑ์ในการประเมินประสิทธิภาพของ CAIA และแบ่งชุดข้อมูลในการทดลองด้วย 10 fold cross validation ผลของการทดลองในตารางที่ 4.9 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ได้ค่าความถูกต้องที่สูงกว่าชุดข้อมูลดิบ (Original-Dataset) ในทุกๆ ตัวจำแนกประเภทข้อมูลคือ J48 และ RBF ได้ค่าความถูกต้องที่ 98.67% เท่ากัน ส่วน MLP และ NB ได้ค่าความถูกต้องที่ 98.00% และได้ค่าเฉลี่ยของจำนวนช่วงข้อมูลเท่ากับ 3

ตารางที่ 4.9 ผลการทดลองของชุดข้อมูล Iris

| Dataset | Iris | | | |
|---|---|---|---|---|
| | J48 | RBF | MLP | NB |
| Original-Dataset | 96.00% | 95.33% | 97.33% | 96.00% |
| CAIA | 98.67% | 98.67% | 98.00% | 98.00% |

**4.2.2 การทดลองชุดข้อมูล Breast Cancer**

**ขั้นตอนที่ 1 การเตรียมข้อมูลในการทดลอง**

1.1 นำชุดข้อมูล Breast Cancer ที่ดาวน์โหลดมาจาก UCI database มาทำความสะอาดข้อมูลโดยใช้วิธีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้าง สำหรับชุดข้อมูล Breast Cancer เป็นชุดข้อมูลที่มีจำนวนแถวข้อมูล 699 แถวข้อมูล มีจำนวน 10 แอตทริบิวต์ มีคลาสจำนวน 2 คลาส

จากตารางที่ 4.10 ตัวอย่างชุดข้อมูลของ Breast Cancer มีค่าข้อมูลที่สูญหายคือในแอตทริบิวต์ของ X7 ในแถวข้อมูลที่ 8 และ X10 ในแถวข้อมูลที่ 3 ดังนั้นก็ต้องมีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้างด้วยสมการที่ (2.1) ก็คือ ค่าข้อมูลก่อนหน้าสูญหาย + ค่าข้อมูลหลังสูญหาย แล้วหารด้วย 2 สำหรับแอตทริบิวต์ที่ X7 ก็จะได้ คือ (7 + 7) / 2 = 7 และสำหรับแอตทริบิวต์ที่ X10 ก็จะได้คือ (1 + 1) / 2 = 1 ดังแสดงในตารางที่ 4.11

ตารางที่ 4.10 ตัวอย่างข้อมูลที่มีค่าสูญหายของชุดข้อมูล Breast Cancer

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | - | benign |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1054590 | 7 | 3 | 2 | 10 | 5 | 10 | 5 | 4 | 4 | malignant |
| 1054593 | 10 | 5 | 5 | 3 | 6 | 7 | 7 | 10 | 1 | malignant |
| 1057013 | 8 | 4 | 5 | 1 | 2 | - | 7 | 3 | 1 | malignant |
| 1065726 | 5 | 2 | 3 | 4 | 2 | 7 | 3 | 6 | 1 | malignant |
| 1072179 | 10 | 7 | 7 | 3 | 8 | 5 | 7 | 4 | 3 | malignant |

ตารางที่ 4.11 ตัวอย่างข้อมูลที่มีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้างของชุดข้อมูล Breast Cancer

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | Class |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000025 | 5 | 1 | 1 | 1 | 2 | 1 | 3 | 1 | 1 | benign |
| 1002945 | 5 | 4 | 4 | 5 | 7 | 10 | 3 | 2 | 1 | benign |
| 1015425 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 1 | **1** | benign |
| 1016277 | 6 | 8 | 8 | 1 | 3 | 4 | 3 | 7 | 1 | benign |
| 1017023 | 4 | 1 | 1 | 3 | 2 | 1 | 3 | 1 | 1 | benign |
| 1054590 | 7 | 3 | 2 | 10 | 5 | 10 | 5 | 4 | 4 | malignant |
| 1054593 | 10 | 5 | 5 | 3 | 6 | 7 | 7 | 10 | 1 | malignant |
| 1057013 | 8 | 4 | 5 | 1 | 2 | **7** | 7 | 3 | 1 | malignant |
| 1065726 | 5 | 2 | 3 | 4 | 2 | 7 | 3 | 6 | 1 | malignant |
| 1072179 | 10 | 7 | 7 | 3 | 8 | 5 | 7 | 4 | 3 | malignant |

**ขั้นตอนที่ 2 การแบ่งช่วงข้อมูลโดยใช้ CAIA**

2.1 นำชุดข้อมูล Breast Cancer ที่ได้จากขั้นตอนการเตรียมข้อมูลมาหาค่าน้อยที่สุดกำหนดเป็น ($d_1$) และค่าน้อยที่สุดกำหนดเป็น ($d_n$) แล้วเรียงลำดับข้อมูลจากน้อยไปหามาก

2.2 จัดกลุ่มข้อมูลของค่าข้อมูลที่มีคลาสเดียวกันที่อยู่ติดกัน แล้วหาค่ากลางของแต่ละกลุ่มข้อมูลที่อยู่ติดกันโดยใช้สมการที่ (3.1)

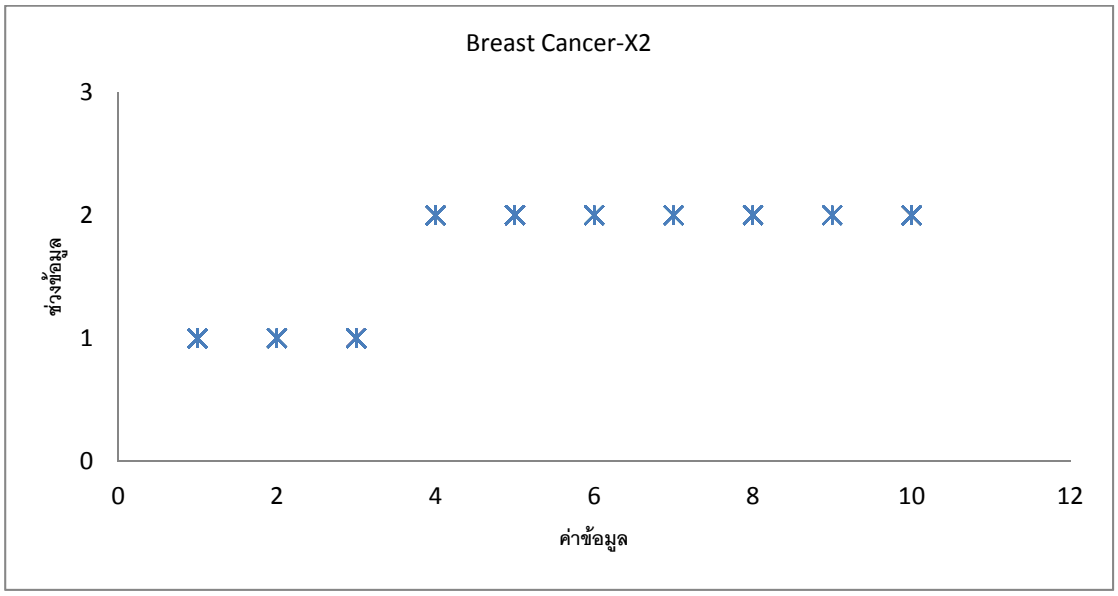
2.3 เซตข้อมูลให้อยู่รูปของตาราง 2D Quanta matrix เพื่อใช้หาค่าความสัมพันธ์ที่กระจายตัวในแต่ละช่วงข้อมูลระหว่างคลาสและแอตทริบิวต์

2.4 หาค่าการกระจายตัวของข้อมูลในแต่ละช่วงข้อมูลโดยใช้สมการที่ (3.2) แล้วหาช่วงข้อมูลที่มีค่าการกระจายตัวน้อยที่สุดเพื่อใช้ในการหลอมรวมกันกับช่วงข้อมูลที่อยู่ติดกัน

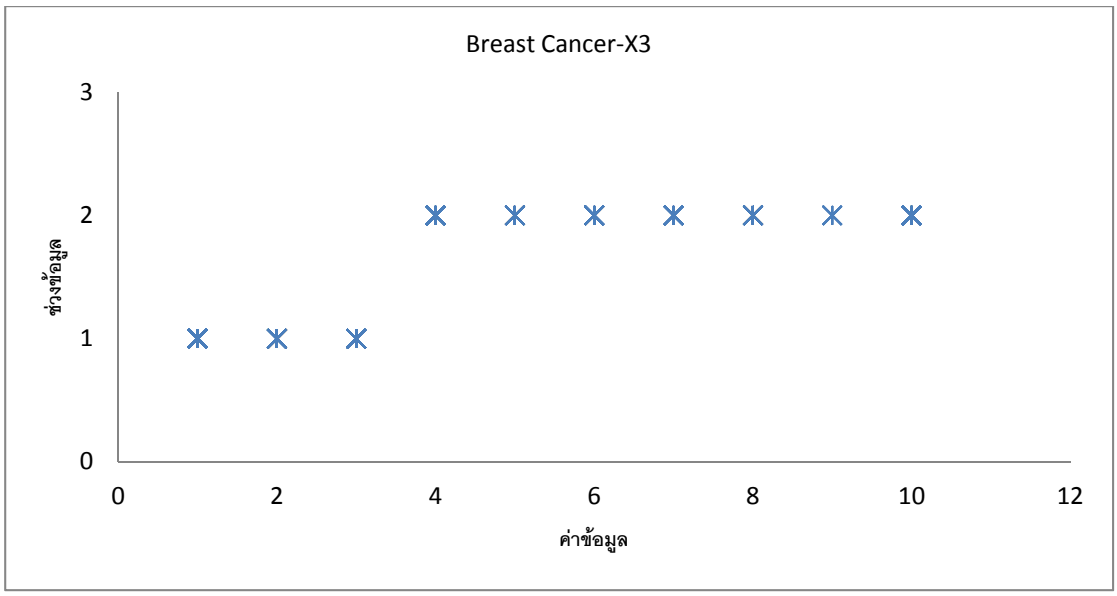
2.5 ทดสอบการหลอมรวมกันกับช่วงข้อมูลที่อยู่ทางซ้ายและทางขวา แล้วหาค่าเฉลี่ยของค่าการกระจายตัวทั้งช่วงข้อมูลอยู่ทางซ้ายและช่วงข้อมูลที่อยู่ทางขวาด้วยสมการที่ (3.3) และเปรียบเทียบค่าเฉลี่ยที่ได้ ถ้าค่าเฉลี่ยของช่วงข้อมูลใดที่มีการหลอมรวมกันแล้วมีค่ามากกว่า ก็ให้เลือกหลอมรวมกับช่วงข้อมูลนั้น จึงกว่าจำนวนของช่วงข้อมูลที่ได้จะเท่ากับจำนวนของคลาสนั้นก็คือ 2

ผลของการแบ่งช่วงข้อมูลชุดข้อมูล Breast Cancer สามารถแบ่งตามแอตทริบิวต์ได้ 9 แอตทริบิวต์คือ X2 = Clump Thickness X3 = Uniformity of Cell Size X4 = Uniformity of Cell Shape X5 = Marginal Adhesion X6 =Single Epithelial Cell Size X7 = Bare Nuclei X8 = Bland Chromatin X9 = Normal Nucleoli และ X10 = Mitoses มีรายละเอียดดังภาพประกอบ 4.6-4.14 และสำหรับ X1 = Sample code number ไม่มีการแบ่งช่วงข้อมูลเนื่องจากว่าเป็นแอตทริบิวต์ที่ใช้ในการนับจำนวนข้อมูลเท่านั้น

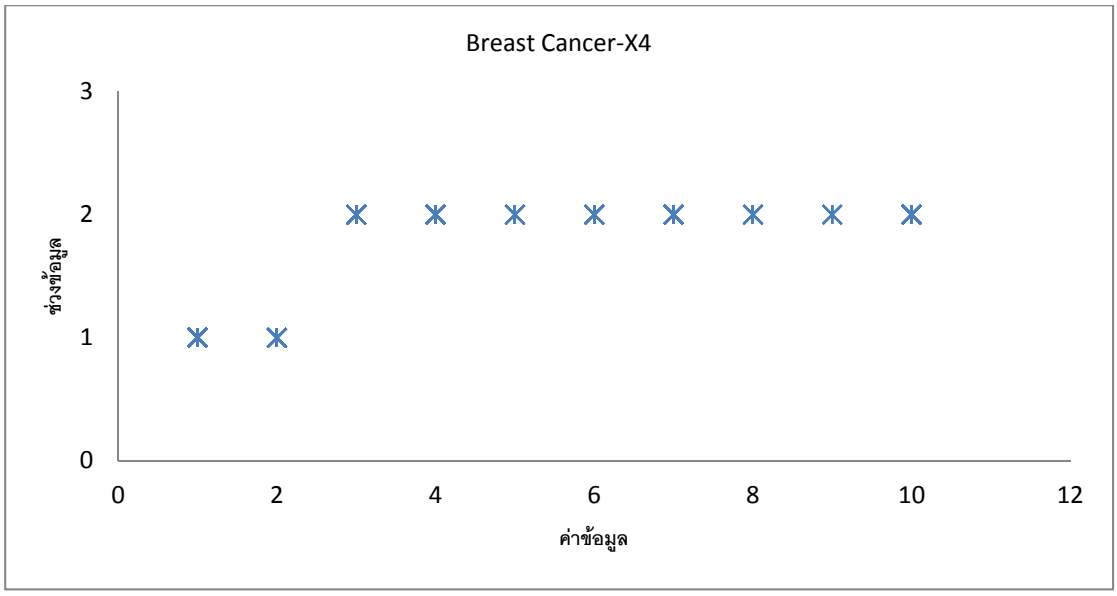
จากภาพประกอบ 4.6 เป็นการแสดงผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลโดยใช้วิธีการของ CAIA แนวตั้งเป็นจำนวนช่วงข้อมูล และแนวนอนเป็นจำนวนของแถวข้อมูล โดยที่เส้นทึบจะเป็นข้อมูลดิบจากแอตทริบิวต์ X2 ของชุดข้อมูล Breast Cancer และเส้นประเป็นข้อมูลที่ผ่านการแบ่งช่วงข้อมูล ดังนั้นจำนวนช่วงข้อมูลที่ได้คือ 2 ช่วงข้อมูล



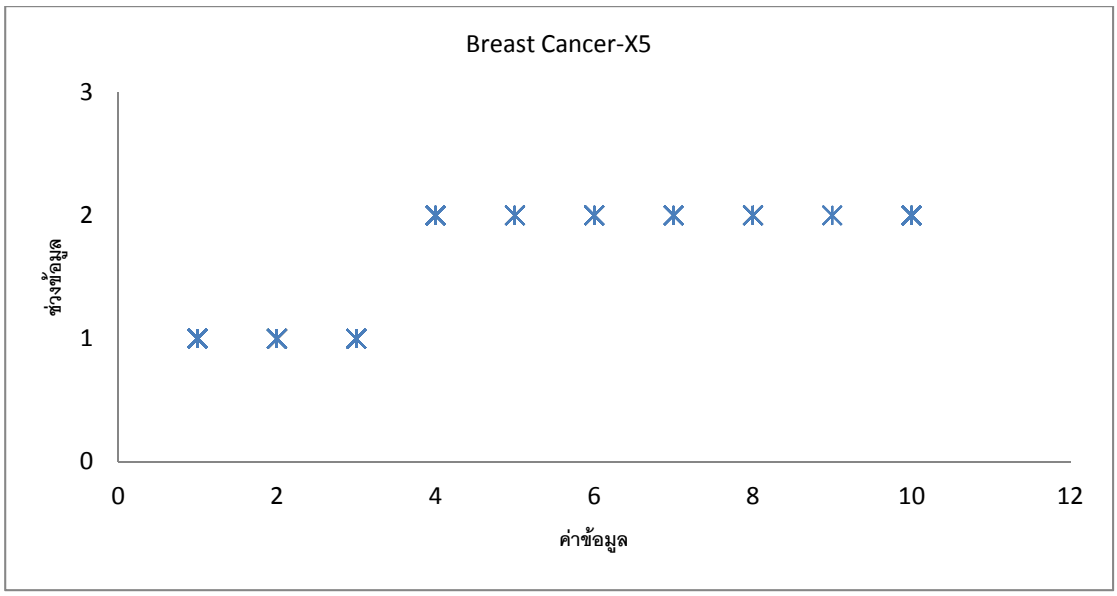

ภาพประกอบ 4.6 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X2 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA



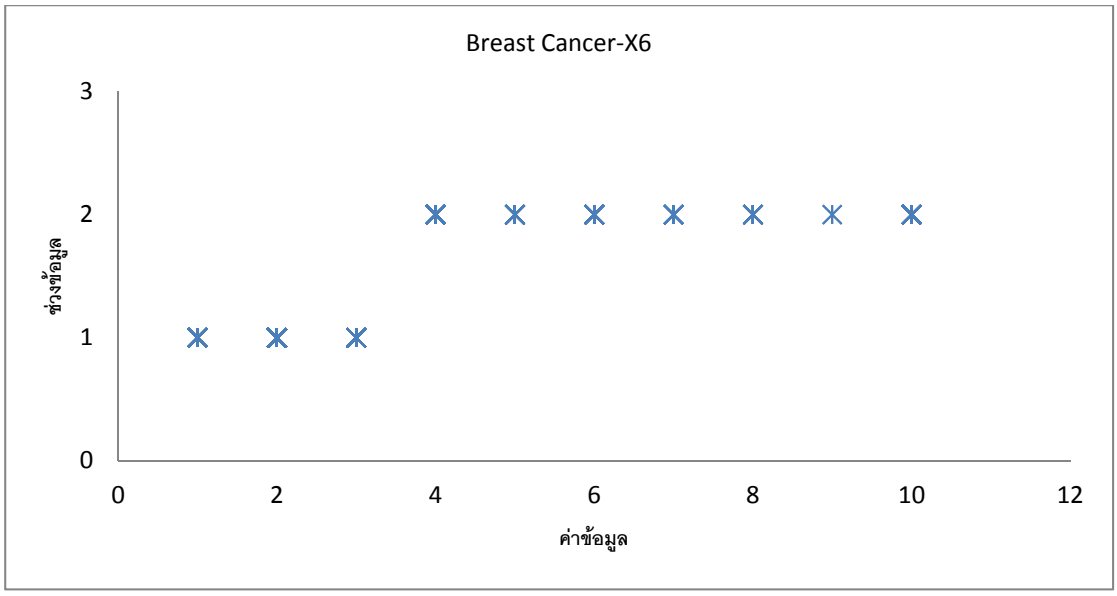

ภาพประกอบ 4.7 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X3 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA



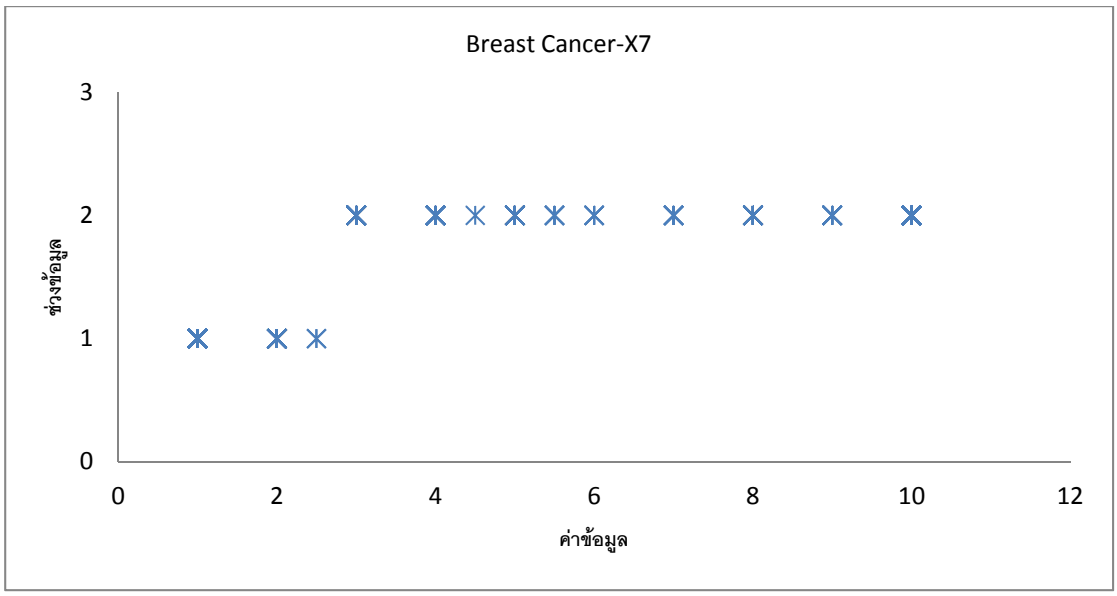

ภาพประกอบ 4.8 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X4 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA



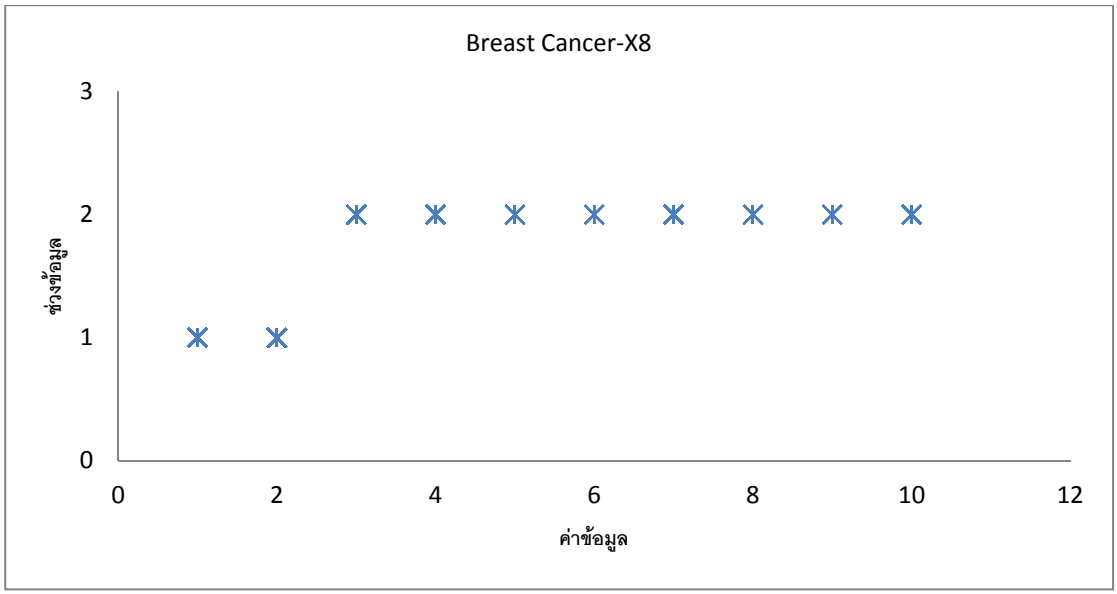

ภาพประกอบ 4.9 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X5 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA



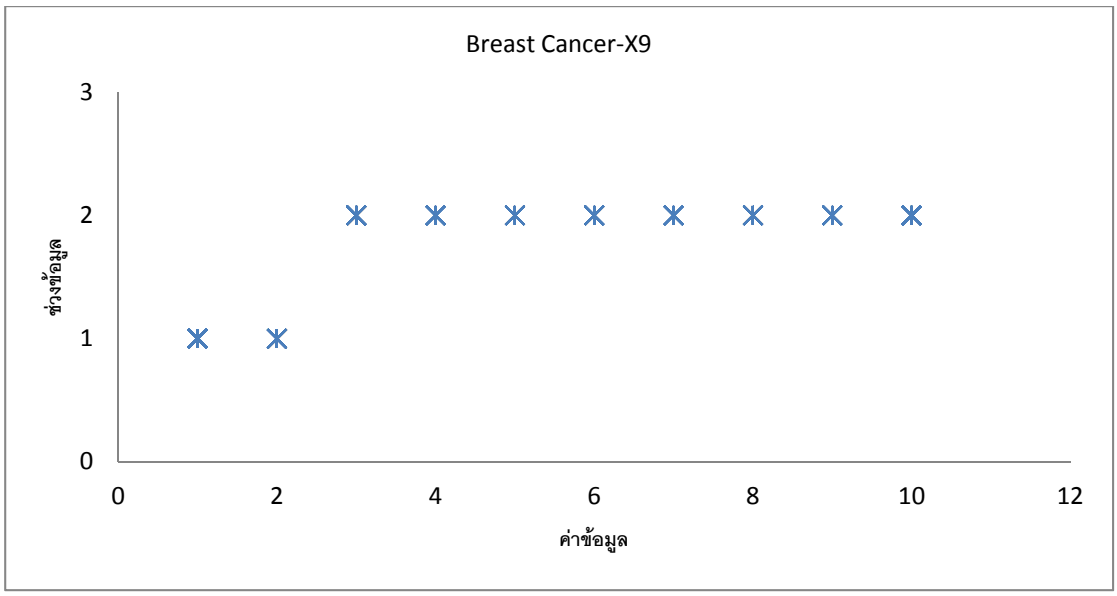

ภาพประกอบ 4.10 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X6 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA



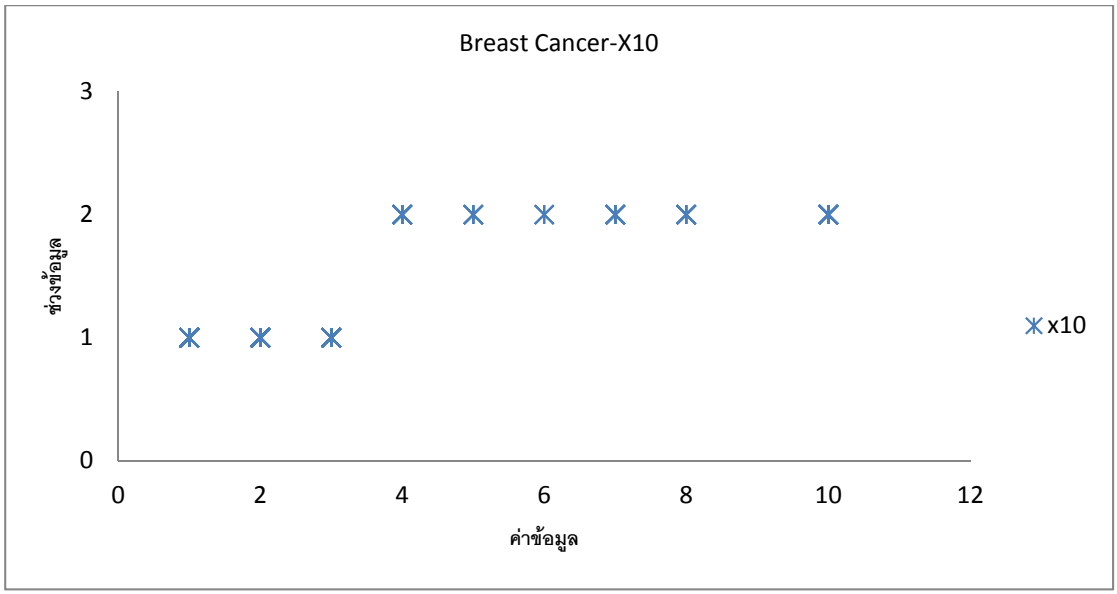

ภาพประกอบ 4.11 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X7 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA




ภาพประกอบ 4.12 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X8 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA




ภาพประกอบ 4.13 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X9 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA




ภาพประกอบ 4.14 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X10 ของชุดข้อมูล Breast Cancer โดยใช้ CAIA

**ขั้นตอนที่ 3 การประเมินประสิทธิภาพในการทดลองของ CAIA**

นำชุดข้อมูล Breast Cancer ที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ไปทดสอบประสิทธิภาพกับตัวจำแนกประเภทในโปรแกรม WEKA คือ J48 RBF MLP และ NB โดยใช้ค่าความถูกต้องเป็นเกณฑ์ในการประเมินประสิทธิภาพของ CAIA และแบ่งชุดข้อมูลในการทดลองด้วย 10 fold cross validation ผลของการทดลองในตารางที่ 4.12 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ได้ค่าความถูกต้องที่สูงกว่าชุดข้อมูลดิบในทุกๆ ตัวจำแนกประเภทข้อมูลทั้ง J48 MLP และ NB คือ 98.67% 95.42% และ 96.57% ยกเว้นตัวจำแนกประเภทข้อมูล RBF ที่ได้ค่าความถูกต้องที่เท่ากันกับชุดข้อมูลดิบ และได้ค่าเฉลี่ยของจำนวนช่วงข้อมูลเท่ากับ 2

ตารางที่ 4.12 ผลการทดลองของชุดข้อมูล Breast Cancer

| Dataset | Breast Cancer | | | |
|---|---|---|---|---|
| | J48 | RBF | MLP | NB |
| Original-Dataset | 93.85% | 95.99% | 95.14% | 96.14% |
| CAIA | 98.67% | 95.99% | 95.42% | 96.57% |

**4.2.3 การทดลองชุดข้อมูล Heart Disease**

**ขั้นตอนที่ 1 การเตรียมข้อมูลในการทดลอง**

1.1 นำชุดข้อมูล Heart Disease ที่ดาวน์โหลดมาจาก UCI database มาทำความสะอาดข้อมูลโดยใช้วิธีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้าง สำหรับชุดข้อมูล Heart Disease เป็นชุดข้อมูลที่มีจำนวนแถวข้อมูล 303 แถวข้อมูล มีจำนวน 13 แอตทริบิวต์ มีคลาสจำนวน 5 คลาส

จากตารางที่ 4.13 ตัวอย่างชุดข้อมูลของ Heart Disease มีค่าข้อมูลที่สูญหายคือ ในแอตทริบิวต์ของ X3 ในแถวข้อมูลที่ 7 และ X13 ในแถวข้อมูลที่ 4 ดังนั้นก็ต้องมีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้างด้วยสมการที่ (2.1) สำหรับแอตทริบิวต์ที่ X3 ก็จะได้ คือ $(4 + 4) / 2 = 4$ และสำหรับแอตทริบิวต์ที่ X13 ก็จะได้คือ $(3 + 3) / 2 = 3$ ดังแสดงในตารางที่ 4.14

ตารางที่ 4.13 ตัวอย่างข้อมูลที่มีค่าสูญหายของชุดข้อมูล Heart Disease

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 3 | 0 | 6 | 0 |
| 37 | 1 | 3 | 130 | 250 | 0 | 0 | 187 | 0 | 3.5 | 3 | 0 | 3 | 0 |
| 41 | 0 | 2 | 130 | 204 | 0 | 2 | 172 | 0 | 1.4 | 1 | 0 | 3 | 0 |
| 56 | 1 | 2 | 120 | 236 | 0 | 0 | 178 | 0 | 0.8 | 1 | 0 | - | 0 |
| 57 | 0 | 4 | 120 | 354 | 0 | 0 | 163 | 1 | 0.6 | 1 | 0 | 3 | 0 |
| 61 | 0 | 4 | 130 | 330 | 0 | 2 | 169 | 0 | 0 | 1 | 0 | 3 | 1 |
| 60 | 1 | - | 130 | 253 | 0 | 0 | 144 | 1 | 1.4 | 1 | 1 | 7 | 1 |
| 54 | 1 | 4 | 124 | 266 | 0 | 2 | 109 | 1 | 2.2 | 2 | 1 | 7 | 1 |
| 50 | 1 | 3 | 140 | 233 | 0 | 0 | 163 | 0 | 0.6 | 2 | 1 | 7 | 1 |
| 41 | 1 | 4 | 110 | 172 | 0 | 2 | 158 | 0 | 0 | 1 | 0 | 7 | 1 |

ตารางที่ 4.14 ตัวอย่างข้อมูลที่มีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้างของชุดข้อมูล Heart Disease

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | X11 | X12 | X13 | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 3 | 0 | 6 | 0 |
| 37 | 1 | 3 | 130 | 250 | 0 | 0 | 187 | 0 | 3.5 | 3 | 0 | 3 | 0 |
| 41 | 0 | 2 | 130 | 204 | 0 | 2 | 172 | 0 | 1.4 | 1 | 0 | 3 | 0 |
| 56 | 1 | 2 | 120 | 236 | 0 | 0 | 178 | 0 | 0.8 | 1 | 0 | **3** | 0 |
| 57 | 0 | 4 | 120 | 354 | 0 | 0 | 163 | 1 | 0.6 | 1 | 0 | 3 | 0 |
| 61 | 0 | 4 | 130 | 330 | 0 | 2 | 169 | 0 | 0 | 1 | 0 | 3 | 1 |
| 60 | 1 | **4** | 130 | 253 | 0 | 0 | 144 | 1 | 1.4 | 1 | 1 | 7 | 1 |
| 54 | 1 | 4 | 124 | 266 | 0 | 2 | 109 | 1 | 2.2 | 2 | 1 | 7 | 1 |
| 50 | 1 | 3 | 140 | 233 | 0 | 0 | 163 | 0 | 0.6 | 2 | 1 | 7 | 1 |
| 41 | 1 | 4 | 110 | 172 | 0 | 2 | 158 | 0 | 0 | 1 | 0 | 7 | 1 |

**ขั้นตอนที่ 2 การแบ่งช่วงข้อมูลโดยใช้ CAIA**

2.1 นำชุดข้อมูล Heart Disease ที่ได้จากขั้นตอนการเตรียมข้อมูลมาหาค่าน้อยที่สุดกำหนดเป็น $(d_1)$ และค่าน้อยที่สุดกำหนดเป็น $(d_n)$ แล้วเรียงลำดับข้อมูลจากน้อยไปหามาก

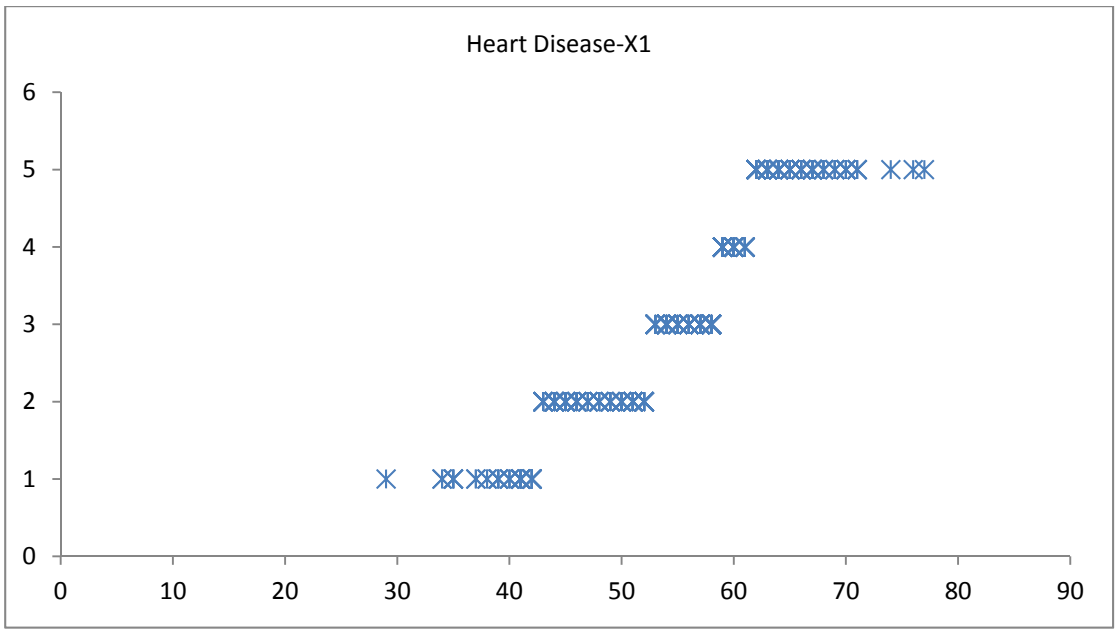
2.2 จัดกลุ่มข้อมูลของค่าข้อมูลที่มีคลาสเดียวกันที่อยู่ติดกัน แล้วหาค่ากลางของแต่ละกลุ่มข้อมูลที่อยู่ติดกันโดยใช้สมการที่ (3.1)

2.3 เซตข้อมูลให้อยู่รูปของตาราง 2D Quanta Matrix เพื่อใช้หาค่าความสัมพันธ์ที่กระจายตัวในแต่ละช่วงข้อมูลระหว่างคลาสและแอตทริบิวต์

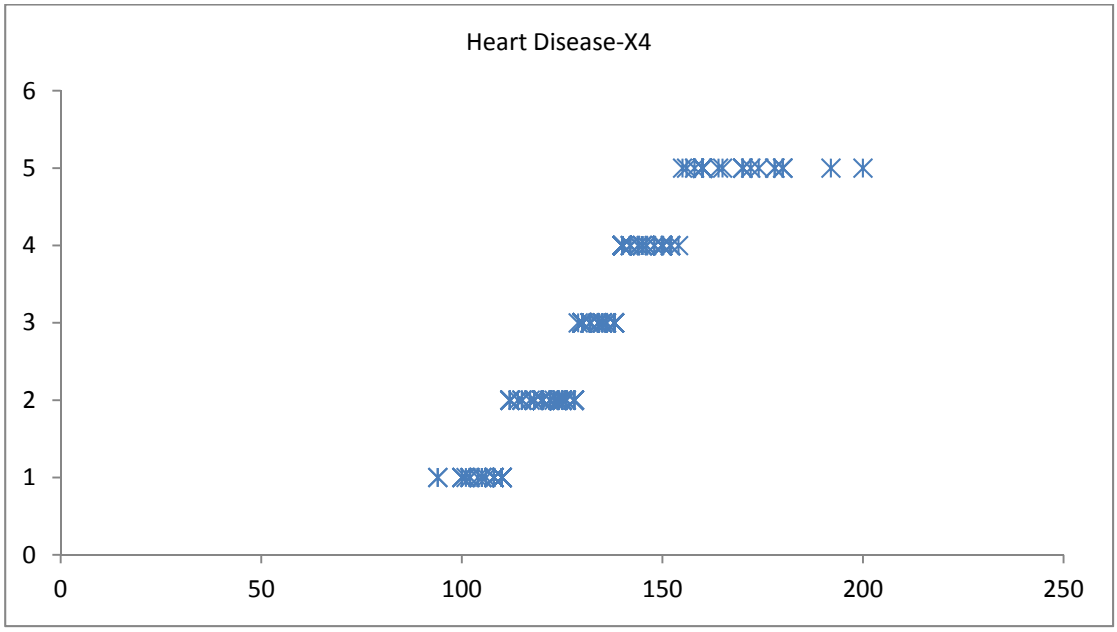
2.4 หาค่าการกระจายตัวของข้อมูลในแต่ละช่วงข้อมูลโดยใช้สมการที่ (3.2) แล้วหาช่วงข้อมูลที่มีค่าการกระจายตัวน้อยที่สุดเพื่อใช้ในการหลอมรวมกันกับช่วงข้อมูลที่อยู่ติดกัน

2.5 ทดสอบการหลอมรวมกันกับช่วงข้อมูลที่อยู่ทางซ้ายและทางขวา แล้วหาค่าเฉลี่ยของค่าการกระจายตัวทั้งช่วงข้อมูลอยู่ทางซ้ายและช่วงข้อมูลที่อยู่ทางขวาด้วยสมการที่ (3.3) และเปรียบเทียบค่าเฉลี่ยที่ได้ ถ้าค่าเฉลี่ยของช่วงข้อมูลใดที่มีการหลอมรวมกันแล้วมีค่ามากกว่า ก็ให้เลือกหลอมรวมกับช่วงข้อมูลนั้น จนกว่าจำนวนของช่วงข้อมูลที่ได้จะเท่ากับจำนวนของคลาสนั้นก็คือ 5

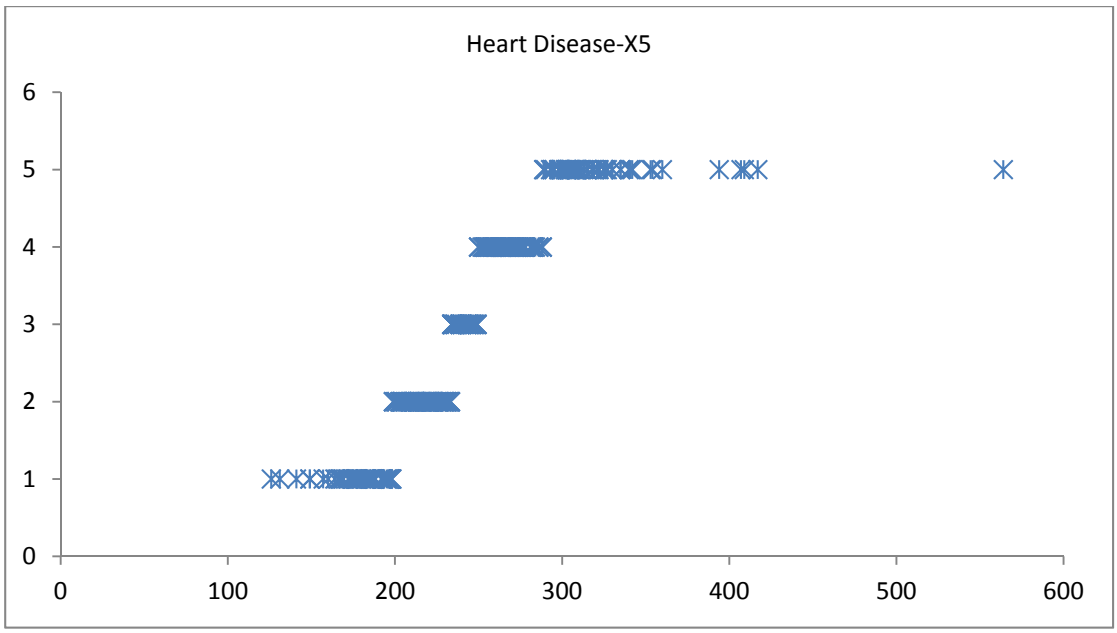
ผลของการแบ่งช่วงข้อมูลชุดข้อมูล Heart Disease สามารถแบ่งตามแอตทริบิวต์ได้ 5 แอตทริบิวต์คือ X1 = age X4 = trestbps X5 = chol X8 = thalach และ X10 = oldpeak มีรายละเอียดดังภาพประกอบ 4.15-4.19 และสำหรับ X2 = sex X3= cp X6 = fbs X7 = restecg X9 = exang X11 = slope X12 = ca และ X13 = thal ไม่มีการแบ่งช่วงข้อมูลเนื่องจากค่าข้อมูลที่แตกต่างกันในแอตทริบิวต์มีน้อยกว่าจำนวนของคลาสนั้นเอง



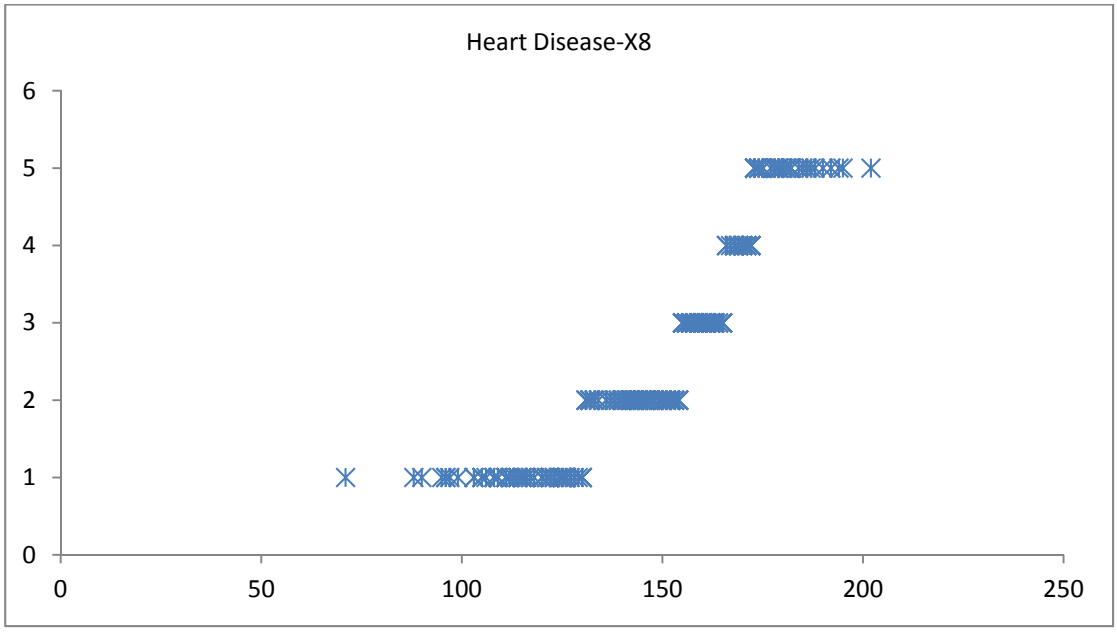

ภาพประกอบ 4.15 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X1 ของชุดข้อมูล Heart Disease โดยใช้ CAIA



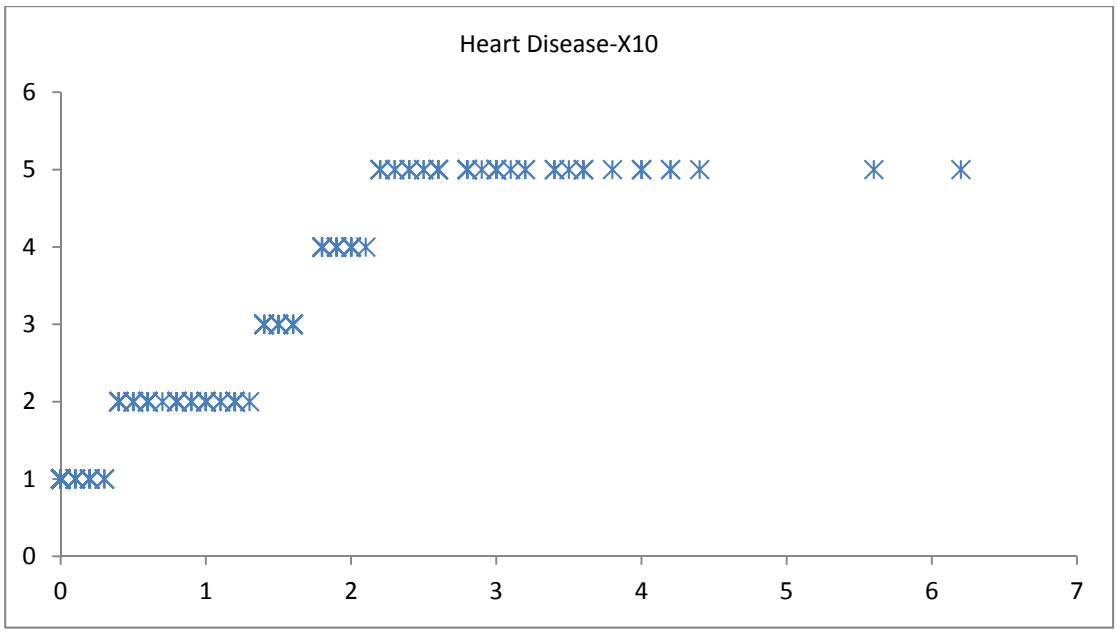

ภาพประกอบ 4.16 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X4 ของชุดข้อมูล Heart Disease โดยใช้ CAIA




ภาพประกอบ 4.17 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X5 ของชุดข้อมูล Heart Disease โดยใช้ CAIA




ภาพประกอบ 4.18 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X8 ของชุดข้อมูล Heart Disease โดยใช้ CAIA




ภาพประกอบ 4.19 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X10 ของชุดข้อมูล Heart Disease โดยใช้ CAIA

**ขั้นตอนที่ 3 การประเมินประสิทธิภาพในการทดลองของ CAIA**

นำชุดข้อมูล Heart Disease ที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ไปทดสอบประสิทธิภาพกับตัวจำแนกประเภทในโปรแกรม WEKA คือ J48 RBF MLP และ NB โดยใช้ค่าความถูกต้องเป็นเกณฑ์ในการประเมินประสิทธิภาพของ CAIA และแบ่งชุดข้อมูลในการทดลองด้วย 10 Fold Cross Validation ผลของการทดลองในตารางที่ 4.15 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ได้ค่าความถูกต้องที่สูงกว่าชุดข้อมูลดิบในทุกๆ ตัวจำแนกประเภทข้อมูลคือ J48 RBF MLP และ NB ได้ค่าความถูกต้องคือ 90.09% 82.50% 87.45% และ 80.85% ตามลำดับ และได้ค่าเฉลี่ยของจำนวนช่วงข้อมูลคือ 5

ตารางที่ 4.15 ผลการทดลองของชุดข้อมูล Heart Disease

| Dataset | Heart Disease | | | |
|---|---|---|---|---|
| | J48 | RBF | MLP | NB |
| Original-Dataset | 54.79% | 54.13% | 52.15% | 58.75% |
| CAIA | 90.09% | 82.50% | 87.45% | 80.85% |

**4.2.4 การทดลองชุดข้อมูล Glass**

**ขั้นตอนที่ 1 การเตรียมข้อมูลในการทดลอง**

1.1 นำชุดข้อมูล Glass ที่ดาวน์โหลดมาจาก UCI database มาทำความสะอาดข้อมูล สำหรับชุดข้อมูล Glass เป็นชุดข้อมูลที่มีจำนวนแถวข้อมูล 214 แถวข้อมูล มีจำนวน 9 แอตทริบิวต์ มีคลาสจำนวน 7 คลาส และไม่มีข้อมูลที่สูญหาย ดังนั้นจึงไม่จำเป็นต้องมีการแทนค่าข้อมูลที่สูญหาย

**ขั้นตอนที่ 2 การแบ่งช่วงข้อมูลโดยใช้ CAIA**

2.1 นำชุดข้อมูล Glass ที่ได้จากขั้นตอนการเตรียมข้อมูลมาหาค่าน้อยที่สุดกำหนดเป็น ($d_1$) และค่าน้อยที่สุดกำหนดเป็น ($d_n$) แล้วเรียงลำดับข้อมูลจากน้อยไปหามาก

2.2 จัดกลุ่มข้อมูลของค่าข้อมูลที่มีคลาสเดียวกันที่อยู่ติดกัน แล้วหาค่ากลางของแต่ละกลุ่มข้อมูลที่อยู่ติดกันโดยใช้สมการที่ (3.1)

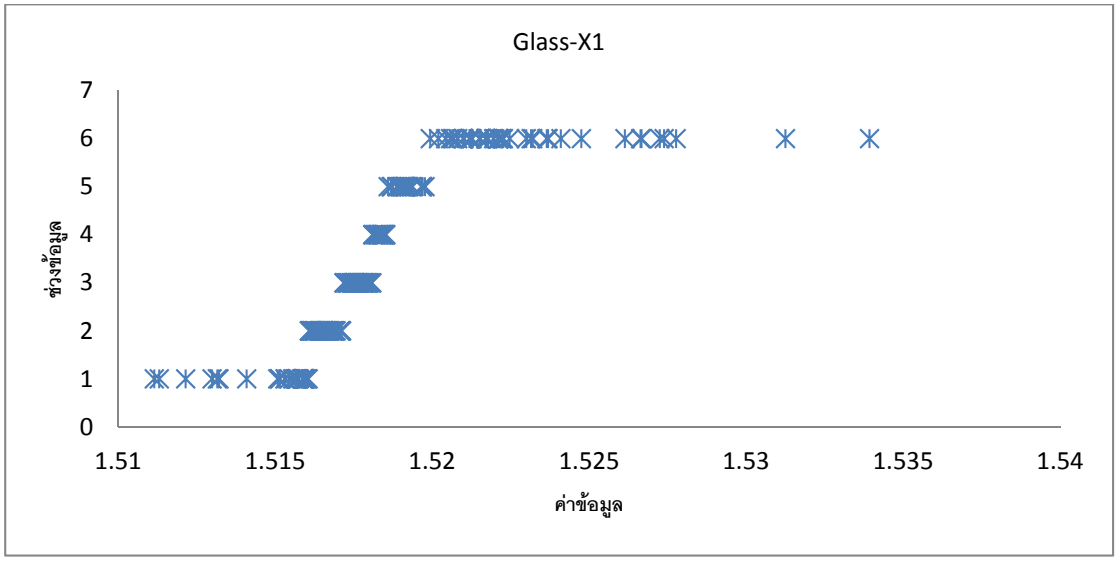
2.3 เซตข้อมูลให้อยู่รูปของตาราง 2D Quanta Matrix เพื่อใช้หาค่าความสัมพันธ์ที่กระจายตัวในแต่ละช่วงข้อมูลระหว่างคลาสและแอตทริบิวต์

2.4 หาค่าการกระจายตัวของข้อมูลในแต่ละช่วงข้อมูลโดยใช้สมการที่ (3.2) แล้วหาช่วงข้อมูลที่มีค่าการกระจายตัวน้อยที่สุดเพื่อใช้ในการหลอมรวมกันกับช่วงข้อมูลที่อยู่ติดกัน

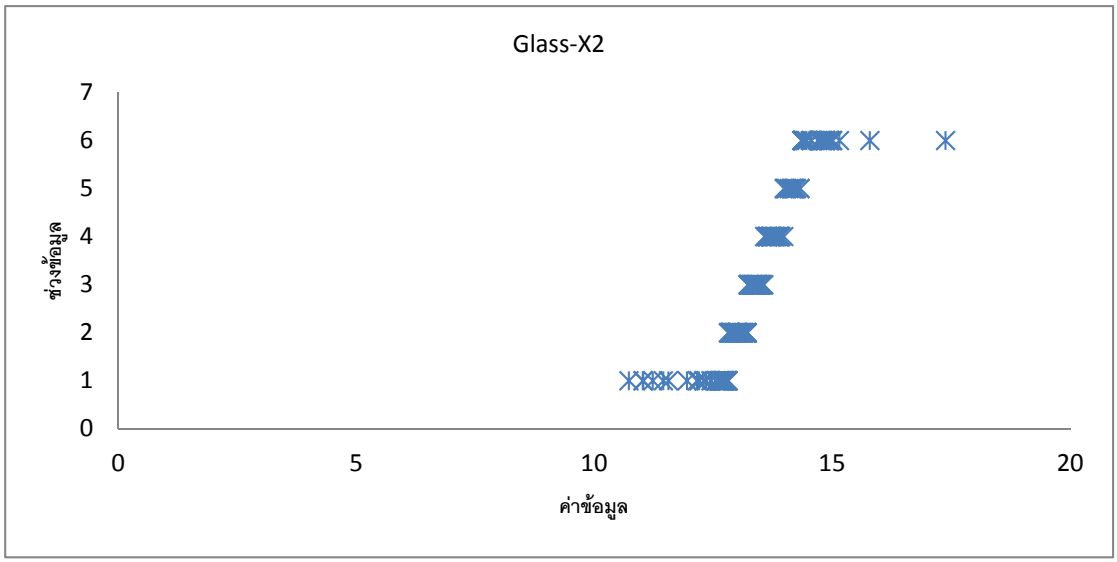
2.5 ทดสอบการหลอมรวมกันกับช่วงข้อมูลที่อยู่ทางซ้ายและทางขวา แล้วหาค่าเฉลี่ยของค่าการกระจายตัวทั้งช่วงข้อมูลอยู่ทางซ้ายและช่วงข้อมูลที่อยู่ทางขวาด้วยสมการที่ (3.3) และเปรียบเทียบค่าเฉลี่ยที่ได้ ถ้าค่าเฉลี่ยของช่วงข้อมูลใดที่มีการหลอมรวมกันแล้วมีค่ามากกว่า ก็ให้เลือกหลอมรวมกับช่วงข้อมูลนั้น จนกว่าจำนวนของช่วงข้อมูลที่ได้จะเท่ากับจำนวนของคลาสนั้นก็คือ 6

ผลของการแบ่งช่วงข้อมูลชุดข้อมูล Glass สามารถแบ่งตามแอตทริบิวต์ได้ 9 แอตทริบิวต์คือ X1 = RI X2 = Na X3 = Mg X4 = Al X5 = Si X6 = K X7 = Ca X8 = Ba และ X9 = Fe มีรายละเอียดดังภาพประกอบ 4.19 ถึง 4.27

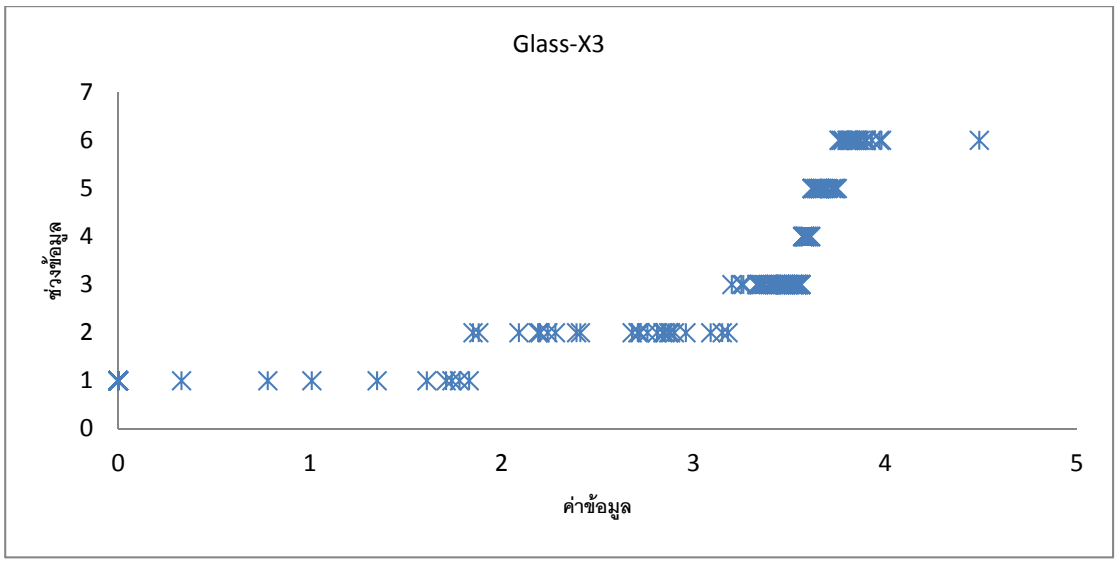
จากภาพประกอบ 4.20 เป็นการแสดงผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลโดยใช้วิธีการของ CAIA แนวตั้งเป็นจำนวนช่วงข้อมูล และแนวนอนเป็นจำนวนของแถวข้อมูล โดยที่เส้นทึบจะเป็นข้อมูลดิบจากแอตทริบิวต์ X1 ของชุดข้อมูล Glass และเส้นประเป็นข้อมูลที่ผ่านการแบ่งช่วงข้อมูล ดังนั้นจำนวนช่วงข้อมูลที่ได้คือ 6 ช่วงข้อมูล



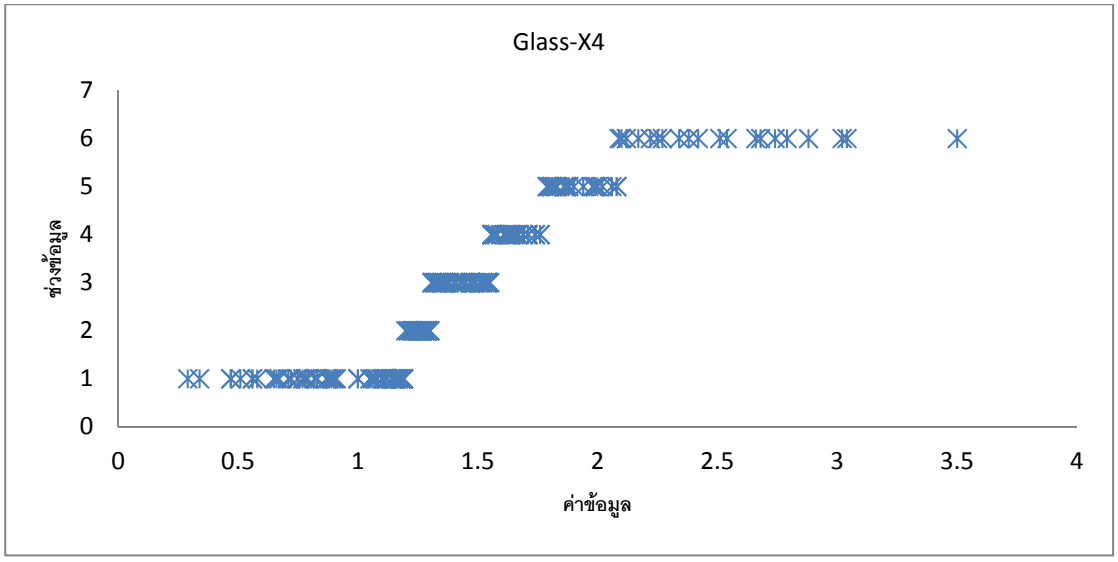

ภาพประกอบ 4.20 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X1 ของชุดข้อมูล Glass โดยใช้ CAIA



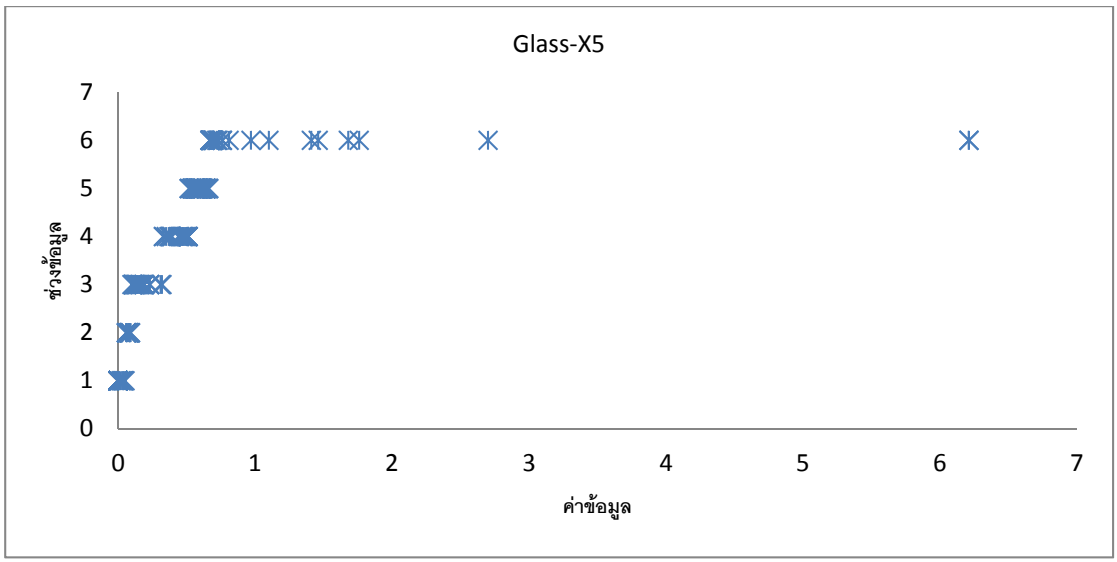

ภาพประกอบ 4.21 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X2 ของชุดข้อมูล Glass โดยใช้ CAIA



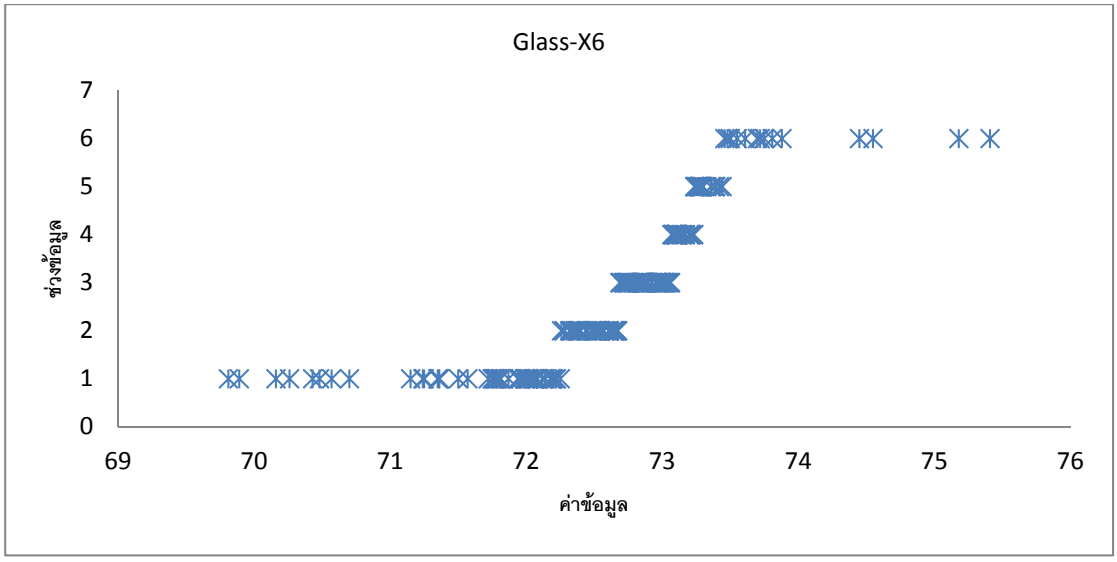

ภาพประกอบ 4.22 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X3 ของชุดข้อมูล Glass โดยใช้ CAIA



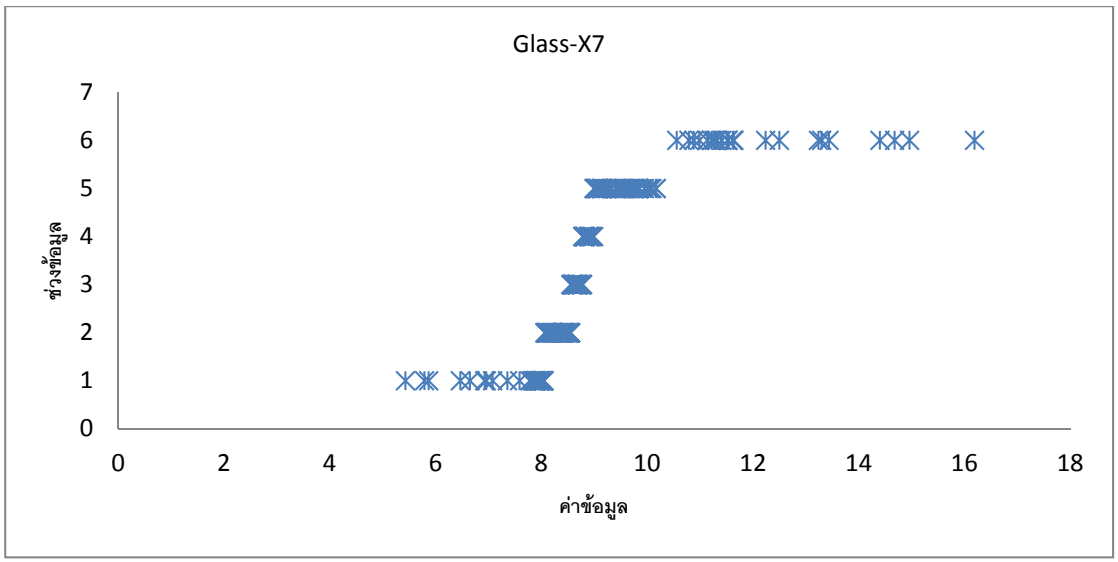

ภาพประกอบ 4.23 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X4 ของชุดข้อมูล Glass โดยใช้ CAIA



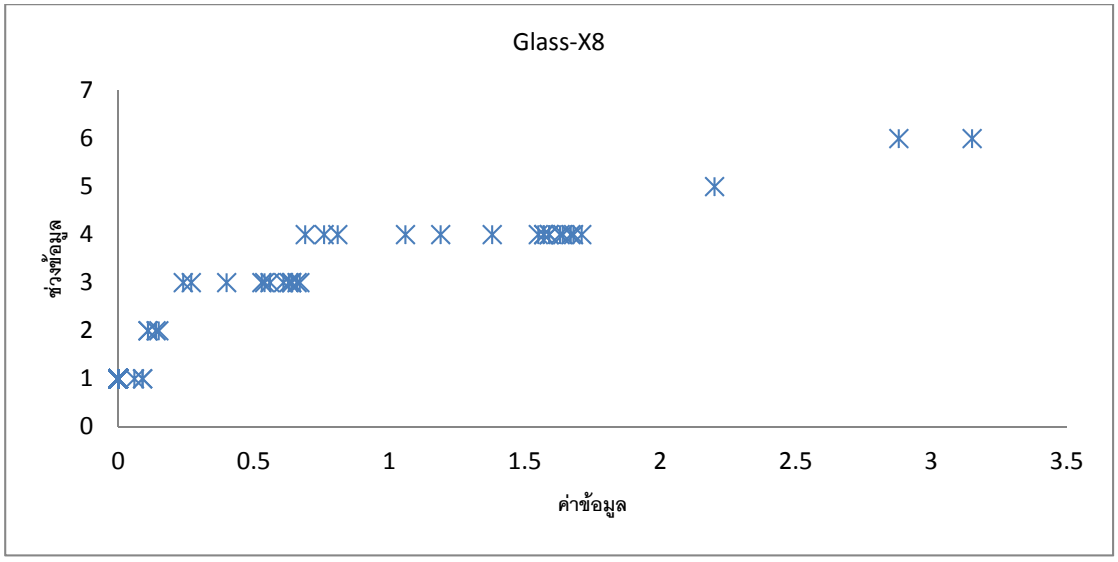

ภาพประกอบ 4.24 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X5 ของชุดข้อมูล Glass โดยใช้ CAIA



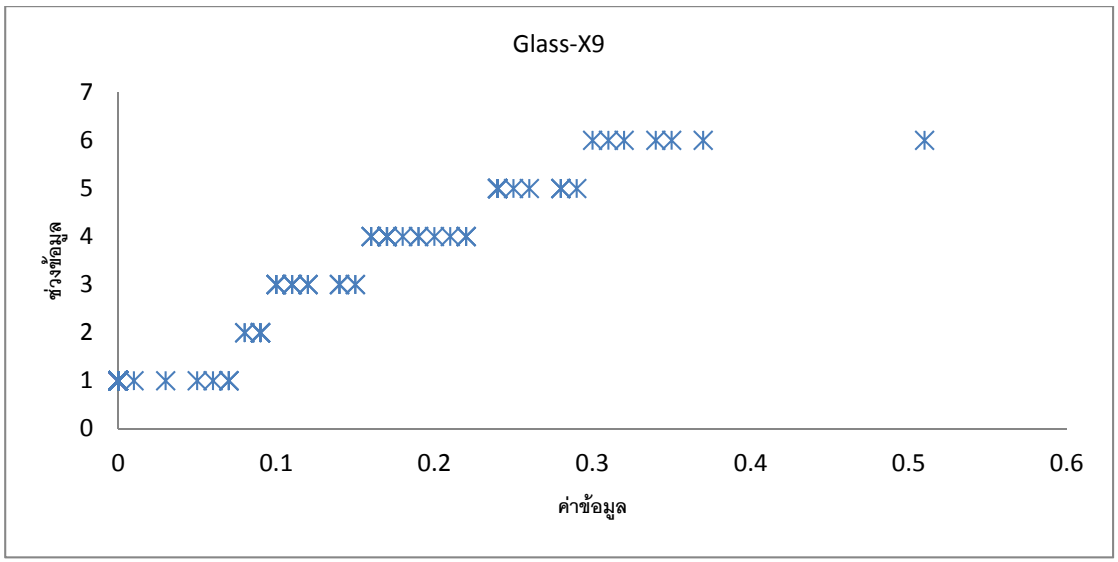

ภาพประกอบ 4.25 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X6 ของชุดข้อมูล Glass โดยใช้ CAIA




ภาพประกอบ 4.26 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X7 ของชุดข้อมูล Glass โดยใช้ CAIA




ภาพประกอบ 4.27 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X8 ของชุดข้อมูล Glass โดยใช้ CAIA




ภาพประกอบ 4.28 ตัวอย่างผลลัพธ์ที่ได้จากการแบ่งช่วงข้อมูลของ X9 ของชุดข้อมูล Glass โดยใช้ CAIA

**ขั้นตอนที่ 3 การประเมินประสิทธิภาพในการทดลองของ CAIA**

นำชุดข้อมูล Glass ที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ไปทดสอบประสิทธิภาพกับตัวจำแนกประเภทในโปรแกรม WEKA คือ J48 RBF MLP และ NB โดยใช้ค่าความถูกต้องเป็นเกณฑ์ในการประเมินประสิทธิภาพของ CAIA และแบ่งชุดข้อมูลในการทดลองด้วย 10 fold cross validation ผลของการทดลองในตารางที่ 4.16 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยวิธีการของ CAIA ได้ค่าความถูกต้องที่สูงกว่าชุดข้อมูลดิบในทุกๆ ตัวจำแนกประเภทข้อมูลคือ J48 RBF MLP และ NB ได้ค่าความถูกต้องคือ 91.55% 69.63% 68.95% และ 61.22% ตามลำดับ และได้ค่าเฉลี่ยของจำนวนช่วงข้อมูลเท่ากับ 6

ตารางที่ 4.16 ผลการทดลองของชุดข้อมูล Glass

| Dataset | Glass | | | |
|---|---|---|---|---|
| | J48 | RBF | MLP | NB |
| Original-Dataset | 66.82% | 64.02% | 67.29% | 48.60% |
| CAIA | 91.55% | 69.63% | 68.95% | 61.22% |

### 4.2.5 ผลการทดลองเปรียบเทียบขั้นตอนวิธีในการแบ่งช่วงข้อมูล

ในการทดลองเปรียบเทียบขั้นตอนวิธีในการแบ่งช่วงข้อมูลของ CAIA ผู้วิจัยได้ทำการเปรียบเทียบกับขั้นตอนวิธีในการแบ่งช่วงข้อมูลที่ได้รับนิยมในปัจจุบันกับ 6 ขั้นตอนวิธีในการแบ่งช่วงข้อมูล โดยมีคุณลักษณะของแต่ละขั้นตอนวิธีดังแสดงในตารางที่ 4.17

ตารางที่ 4.17 คุณลักษณะของแต่ละขั้นตอนวิธีที่ใช้ในการประเมินประสิทธิภาพกับ CAIA

| **Algorithms** | **Characteristics** | | | | | **Discretization Method** |
|---|---|---|---|---|---|---|
| | **Supervised** | **Univariate** | **Merge** | **Global** | **Static** | |
| Equal-width [18] | N | Y | N | Y | Y | binning |
| Equal-frequency [18] | N | Y | N | Y | Y | binning |
| ChiMerge [16] | Y | Y | Y | Y | Y | statistical |
| IEM [13] | Y | Y | N | N | Y | entropy |
| CAIM [5] | Y | Y | N | Y | Y | statistical |
| CACC [20] | Y | Y | N | Y | Y | statistical |
| CAIA Proposed | Y | Y | Y | Y | Y | statistical |

ในการเปรียบเทียบขั้นตอนวิธีในการแบ่งช่วงข้อมูลนั้นจะใช้เกณฑ์ในการวัดค่าความถูกต้องที่สูงกว่าของตัวจำแนกประเภทข้อมูลคือ J48 RBF MLP และ NB และนอกจากนี้ยังใช้จำนวนของจำนวนช่วงข้อมูลที่ได้จากมีการแบ่งช่วงข้อมูลที่น้อยที่สุดเป็นเกณฑ์ในการประเมินประสิทธิภาพของแต่ละขั้นตอนวิธีโดยมีรายละเอียดดังต่อไปนี้

ตารางที่ 4.18 ผลการทดลองหาค่าความถูกต้องโดยใช้ตัวจำแนกประเภทข้อมูล J48

| Algorithm | Iris | Breast cancer | Heart disease | Glass | Mean Rank |
|---|---|---|---|---|---|
| **EW [6]** | 94.70% | 91.30% | 70.30% | 86.10% | 5.5 |
| **EF [6]** | 94.00% | 90.80% | 72.60% | 87.00% | 5.5 |
| **ChiMerge [7]** | 90.70% | 93.00% | 76.50% | 88.50% | 5.25 |
| **IEM [13]** | 94.00% | 93.60% | 75.20% | 89.60% | 4 |
| **CAIM [4]** | 94.00% | 93.80% | 77.10% | 90.60% | 3 |
| **CACC [2]** | 93.50% | 94.10% | 78.60% | 90.90% | 3 |
| **CAIA (proposed)** | **98.67%** | **96.57%** | **90.09%** | **91.55%** | **1** |

ผลการทดลองจากตารางที่ 4.18 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยขั้นตอนวิธีของ CAIA มีอันดับเฉลี่ยของค่าความถูกต้อง (Mean Rank) อยู่ในอันดับที่ 1 โดยได้ค่าความถูกต้องสูงที่สุดในทุกๆ ชุดข้อมูลเมื่อเปรียบเทียบกับขั้นตอนวิธีอื่นๆ เมื่อใช้ตัวจำแนกประเภทข้อมูลของ J48 คือ ชุดข้อมูล Iris ขั้นตอนวิธีของ CAIA ได้ 98.67% สูงกว่า EW ที่ได้อันดับที่สองคือ 94.70% ชุดข้อมูล Breast Canncer ขั้นตอนวิธีของ CAIA ได้ 96.57% สูงกว่า CACC ที่ได้อันดับที่สองคือ 94.10% ชุดข้อมูล Heart Disease ขั้นตอนวิธีของ CAIA ได้ 90.09% สูงกว่า CACC ที่ได้อันดับที่สองคือ 78.60% และชุดข้อมูล Glass ขั้นตอนวิธีของ CAIA ได้ 91.55% สูงกว่า CACC ที่ได้อันดับที่สองคือ 90.90%

ตารางที่ 4.19 ผลการทดลองหาค่าความถูกต้องโดยใช้ตัวจำแนกประเภทข้อมูล RBF

| Algorithm | Iris | Breast cancer | Heart disease | Glass | Mean Rank |
|---|---|---|---|---|---|
| **EW [6]** | 93.33% | 95.00% | 57.76% | 65.42% | 2.75 |
| **EF [6]** | 92.00% | 95.00% | 54.13% | 63.08% | 4.25 |
| **ChiMerge [7]** | 89.76% | 92.00% | 54.87% | 62.00% | 5.5 |
| **IEM [13]** | 94.00% | 80.98% | 58.09% | 65.70% | 3.25 |
| **CAIM [4]** | 92.67% | 93.42% | 50.49% | 59.81% | 5.5 |
| **CACC [2]** | 92.27% | 94.85% | 51.49% | 56.54% | 5.5 |
| **CAIA (proposed)** | **98.67%** | **95.99%** | **82.50%** | **69.63%** | **1** |

ผลการทดลองจากตารางที่ 4.19 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยขั้นตอนวิธีของ CAIA มีอันดับเฉลี่ยของค่าความถูกต้อง (Mean Rank) อยู่ในอันดับที่ 1 โดยได้ค่าความถูกต้องสูงที่สุดในทุกๆ ชุดข้อมูลเมื่อเปรียบเทียบกับขั้นตอนวิธีอื่นๆ เมื่อใช้ตัวจำแนกประเภทข้อมูลของ RBF คือ ชุดข้อมูล Iris ขั้นตอนวิธีของ CAIA ได้ 98.67% สูงกว่า IEM ที่ได้อันดับที่สองคือ 94.00% ชุดข้อมูล Breast Canncer ขั้นตอนวิธีของ CAIA ได้ 95.99% สูงกว่า EW และ EF ที่ได้อันดับที่สองคือ 95.00% ชุดข้อมูล Heart disease ขั้นตอนวิธีของ CAIA ได้ 82.50% สูงกว่า IEM ที่ได้อันดับที่สองคือ 58.09% และชุดข้อมูล Glass ขั้นตอนวิธีของ CAIA ได้ 69.63% สูงกว่า IEM ที่ได้อันดับที่สองคือ 65.70%

ตารางที่ 4.20 ผลการทดลองหาค่าความถูกต้องโดยใช้ตัวจำแนกประเภทข้อมูล MLP

| Algorithm | Iris | Breast cancer | Heart disease | Glass | Mean Rank |
|---|---|---|---|---|---|
| **EW [6]** | 92.67% | 94.57% | 70.59% | 64.42% | 4 |
| **EF [6]** | 91.33% | 94.71% | 52.15% | 63.55% | 5.25 |
| **ChiMerge [7]** | 93.37% | 92.89% | 54.43% | 63.79% | 4.75 |
| **IEM [13]** | 93.33% | 74.69% | 55.45% | 66.17% | 4.5 |
| **CAIM [4]** | 94.67% | 93.56% | 46.85% | **69.16%** | 3.75 |
| **CACC [2]** | 93.97% | 95.14% | 47.85% | 55.61% | 4.5 |
| **CAIA (proposed)** | **98.00%** | **95.42%** | **87.45%** | **68.95%** | **1.25** |

ผลการทดลองจากตารางที่ 4.20 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยขั้นตอนวิธีของ CAIA มีอันดับเฉลี่ยของค่าความถูกต้อง (Mean Rank) อยู่ในอันดับที่ 1 (Mean Rank คือ 1.25) โดยได้ค่าความถูกต้องสูงที่สุดในทุกๆ ชุดข้อมูลเมื่อเปรียบเทียบกับขั้นตอนวิธีอื่นๆ ยกเว้นในชุดข้อมูลของ Glass เมื่อใช้ตัวจำแนกประเภทข้อมูลของ MLP คือ ชุดข้อมูล Iris ขั้นตอนวิธีของ CAIA ได้ 98.00% สูงกว่า CAIM ที่ได้อันดับที่สองคือ 94.67% ชุดข้อมูล Breast Canncer ขั้นตอนวิธีของ CAIA ได้ 95.42% สูงกว่า CACC ที่ได้อันดับที่สองคือ 95.14% ชุดข้อมูล Heart Disease ขั้นตอนวิธีของ CAIA ได้ 87.45% สูงกว่า EW ที่ได้อันดับที่สองคือ 70.59% และชุดข้อมูล Glass ขั้นตอนวิธีของ CAIA ได้ 68.95% เป็นรองแค่ CAIM ที่ได้อันดับที่ 1 คือ 69.16%

ตารางที่ 4.21 ผลการทดลองหาค่าความถูกต้องโดยใช้ตัวจำแนกประเภทข้อมูล NB

| Algorithm | Iris | Breast cancer | Heart disease | Glass | Mean Rank |
|---|---|---|---|---|---|
| **EW [6]** | 93.33% | **97.28%** | 57.76% | 63.55% | 3.75 |
| **EF [6]** | 92.67% | 96.28% | 58.75% | **69.16%** | 3.75 |
| **ChiMerge [7]** | 93.78% | 91.88% | 59.86% | 54.23% | 4.75 |
| **IEM [13]** | 94.00% | 81.68% | 59.74% | 64.30% | 3.5 |
| **CAIM [4]** | 94.00% | 93.99% | 52.46% | 62.15% | 4.5 |
| **CACC [2]** | 93.34% | 95.28% | 54.46% | 59.35% | 5.25 |
| **CAIA (proposed)** | **98.00%** | **96.57%** | **80.85%** | 61.22% | **2.25** |

ผลการทดลองจากตารางที่ 4.21 แสดงให้เห็นว่าชุดข้อมูลที่ผ่านการแบ่งช่วงข้อมูลด้วยขั้นตอนวิธีของ CAIA มีอันดับเฉลี่ยของค่าความถูกต้อง (Mean Rank) อยู่ในอันดับที่ 1 (Mean Rank คือ 2.25) เมื่อใช้ตัวจำแนกประเภทข้อมูลของ MLP โดยได้ค่าความถูกต้องสูงที่สุดในชุดข้อมูลของ Iris คือ 98.00% โดยมี IEM และ CAIM ที่ได้ค่าความถูกต้องสูงเป็นอันดับที่สองคือ 94.00% ชุดข้อมูลของ Heart Disease ขั้นตอนวิธีของ CAIA ได้ค่าความถูกต้องสูงสุดคือ 80.85% โดยมี ChiMerge อยู่ในอันดับที่สองคือ 59.86% สำหรับชุดข้อมูล Breast Cancer ขั้นตอนวิธีของ CAIA อยู่ในอันดับที่สอง โดยได้ค่าความถูกต้องคือ 96.57% เป็นรองแค่ EW ที่ได้ค่าความถูกต้องคือ 97.28% และชุดข้อมูลของ Glass ขั้นตอนวิธีของ CAIA ได้ค่าความถูกต้องอยู่ในอันดับที่ 5 คือ 61.22% โดยมี EF ได้ค่าความถูกต้องคือ 69.16% IEM ได้ค่าความถูกต้องคือ 64.30% EW ได้ค่าความถูกต้องคือ 63.55% และ CAIM ได้ค่าความถูกต้องคือ 62.15%% เป็นอันดับที่ 1 2 3 และ 4 ตามลำดับ แต่ถึงอย่างไรก็ตาม ขั้นตอนวิธีของ CAIA ก็ยังคงได้ค่าเฉลี่ยอันดับอยู่ในอันดับที่ 1

# บทที่ 5

# บทสรุปและข้อเสนอแนะ

## 5.1 สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอขั้นตอนวิธีใหม่ในการแบ่งช่วงข้อมูลเพื่อใช้ในการทำเหมืองข้อมูลให้มีประสิทธิภาพ โดยใช้ค่าเฉลี่ยการกระจายตัวของข้อมูลระหว่างคลาสและแอตทริบิวต์ของทุกๆ ช่วงข้อมูล หรือเรียกว่า Class Attribute Interval Average Discretization Algorithm (CAIA) มีคุณลักษณะของการแบ่งช่วงข้อมูลเป็น Supervised, Univariate, Global, Merge และ Static และใช้หลักการทางสถิติมาช่วยในการค้นหาช่วงข้อมูลที่ดีที่สุดในการแบ่งช่วงข้อมูล การทำงานของขั้นตอนวิธี CAIA แบ่งออกเป็นสองขั้นตอนหลักๆ คือ ขั้นตอนที่ 1 การเรียงลำดับข้อมูลจากน้อยไปหามาก แล้วทำการ Merge หรือจัดกลุ่มข้อมูลที่มีคลาสเดียวกันที่อยู่ติดกันให้อยู่ในช่วงข้อมูลเดียวกัน เพื่อคำนวณหาค่ากลาง (Midpoint) ของแต่ละช่วงข้อมูล กำหนดเป็นขอบเขตข้อมูลในแต่ละช่วงข้อมูลที่อยู่ติดกัน และจัดการข้อมูลให้อยู่ในรูปแบบของ 2D quanta matrix และขั้นตอนที่ 2 ค้นหาช่วงข้อมูลที่มีค่าการกระจายตัวของข้อมูลที่น้อยที่สุด เพื่อกำหนดจุดที่จะใช้ในการหลอมรวมของสองช่วงข้อมูลที่อยู่ติดกัน โดยจะทำการทดสอบการหลอมรวมกับช่วงข้อมูลที่อยู่ทางขวา และทางซ้ายของช่วงข้อมูลที่มีค่าการกระจายตัวที่น้อยที่สุด แล้วเลือกหลอมรวมกับช่วงข้อมูลที่ได้ค่าเฉลี่ย CAIA สูงกว่า และหยุดเมื่อจำนวนของช่วงข้อมูลที่ได้มีค่าเท่ากับจำนวนของคลาส

CAIA ได้เปรียบเทียบขั้นตอนวิธีในการแบ่งช่วงข้อมูลกับ 6 ขั้นตอนวิธีในการแบ่งช่วงข้อมูลคือ 1) EW 2) EF 3) ChiMerge 4) IEM 5) CAIM และ 6) CACC โดยทดสอบกับ 4 ชุดข้อมูลของ Benchmarks จาก UCI Data Set คือ 1) Iris 2) Breast Cancer 3) Heart Disease และ 4) Glass และใช้ 4 ตัวจำแนกประเภทข้อมูลจากโปรแกรม WEKA คือ J48 RBF MLP และ NB ในการประเมินค่าความถูกต้อง มีการทดสอบแบบ 10 Fold Cross Validation และได้มีการเปรียบเทียบจำนวนของช่วงข้อมูลที่ได้จากการแบ่งช่วงข้อมูล

ผลการทดลองของขั้นตอนวิธีในการแบ่งช่วงข้อมูลของ CAIA สำหรับการทำเหมืองข้อมูลสามารถสรุปได้ 2 ประเด็นดังต่อไปนี้

1) ประเด็นเปรียบเทียบค่าความถูกต้องที่ใช้ในการประเมินประสิทธิภาพของขั้นตอนวิธีที่ได้นำเสนอ โดยขั้นตอนวิธีที่นำเสนอนั้น ได้ค่าเฉลี่ยของค่าความถูกต้อง (Mean Rank) อันดับที่ 1 ในทุกๆ ตัวจำแนกประเภทข้อมูลคือ J48 RBF MLP และ NB เมื่อเปรียบเทียบกับทั้ง 6 ขั้นตอนวิธีในการแบ่งช่วงข้อมูล

2) ประเด็นเปรียบเทียบจำนวนของช่วงข้อมูลที่ได้จากการแบ่งช่วงข้อมูล โดยขั้นตอนวิธีที่นำเสนอมีค่าเฉลี่ยของจำนวนช่วงข้อมูลที่น้อยที่สุด เมื่อเปรียบเทียบกับทั้ง 6 ขั้นตอนวิธีในการแบ่งช่วงข้อมูล

## 5.2 ปัญหาและอุปสรรค

5.2.1 เนื่องจากบางชุดข้อมูลที่ใช้ในการทดสอบมีจำนวนของข้อมูลที่มีค่าที่แตกต่างกันจำนวนมาก และมีจำนวนของแอตทริบิวต์ที่มาก อีกทั้งบางชุดข้อมูลยังมีข้อมูลที่สูญหาย จึงจำเป็นต้องมีการทำความสะอาดข้อมูลก่อนที่จะทำการแบ่งช่วงข้อมูล ทำให้ใช้ต้องเวลานานในขั้นตอนของการเตรียมข้อมูล

5.2.2 เนื่องจากชุดข้อมูลที่ใช้มีจำนวนของแอตทริบิวต์ที่มากทำให้การจำแนกประเภทข้อมูลของ MLP ต้องใช้เวลาในการประมวลผลนาน

## 5.3 ข้อเสนอแนะ

5.3.1 ในงานวิจัยนี้ได้ออกแบบและพัฒนาขั้นตอนวิธีที่สามารถใช้ได้กับข้อมูลที่เป็น Continuous Data เท่านั้น ดังนั้นจึงควรพัฒนาและออกแบบขั้นตอนวิธีที่สามารถใช้ได้กับข้อมูลที่เป็น Mixed Mode Data เนื่องจากข้อมูลที่มีอยู่จริงอาจจะเป็นทั้งข้อมูลที่เป็นทั้งแบบต่อเนื่องและแบบไม่ต่อเนื่อง

# บรรณานุกรม


กาญจนา ทองกลิ่น. 2008. แบบจำลองการแก้ปัญหาความกำกวมของคำจากคลังข้อความโดยใช้เทคนิคคำบริบท. วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต. มหาวิทยาลัยสงขลานครินทร์ สงขลา.

ณสิทธิ์ เหล่าเส็น. 2551. แบบจำลองการกรองข้อมูลอากาศที่มีสิ่งรบกวนโดยใช้โครงข่ายประสาทเทียม. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยสงขลานครินทร์ สงขลา.

พรพิมล ณ นคร. 2549. แบบจำลองระบบพยากรณ์อากาศโดยใช้โครงข่ายประสาทเทียม กรณีศึกษากรมอุตุนิยมวิทยา ประเทศไทย. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยสงขลานครินทร์ สงขลา.

สุคนธ์ทิพย์ วงศ์พันธ์. 2551. การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมเพื่อจำแนกพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา. วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต, มหาวิทยาลัยเกษตรศาสตร์ กรุงเทพฯ.

Blake, C. L., and Merz, C.J. 1998. UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA: University of California, Department of Information and Computer Science.

Chantasut, N., Chroenjit, C., and Tenpraser, C. 2004. Predictive Mining of Rainfall Prediction Using artificial Neural Networkes for chao Phraya River. http://www.nectec.or.th/NTJ/NO15/papers/11.pdf. (accessed 03/23/2008)

Chaoqun, Y., Jianping , L., and Enming, D. 2011. A Discretization Algorithm based on Clustering and CAIR Criterion. IEEE 7th International Conference on Natural Computation, pp. 1424-1429.

Dorado, J., Rabunal, J. R., Rivero, D., Santos, A., and Pazos, A. 2002. Automatic Recurrent ANN Rule Extraction with Genetic Programming. IEEE Proceedings of the 2002 International Joint Conferent on Neural Networks. 12-17 May. pp.1552-1557.

Fayyad, U. M., and Irani, K. B. 1993. Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning in Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI). pp. 1022-1029.

Garcia, S., Luengo, J., Saez, J. A., Lopez V., and Herrera, F. 2011. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. IEEE Transactions on Knowledge and Data Engineering. Vol. X, No. Y.

George H. J., Pat L., 1995. Estimating Continuous Distributions in Bayesian Classifiers. In Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, pp. 338-345.

Han, J., and Kamber, M. 2002. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.

Hua, H., and Zhao, H. 2009. A Discretization Algorithm of Continuous Attributes Based on Supervised Clustering. IEEE/CCPR Conference on Pattern Recognition. pp. 1-5.

Kang, Y., Wang, S., Liu, X., Lai, H., Wang, H., and Miao, B. 2006. An ICA-based multivariate discretization algorithm. In Proceeding of the First International Conference on Knowledge Science, Engineering and Management (KSEM). pp.556.562.

Kerber, R. 1992. ChiMerge: Discretization of Numeric Attributes. In Prodeedings of the 9th National Conference on Artificial Intelligence, pp. 123-128.

Kohavi. R., and Provost, F. 1998.Glossary of terms, Machine Learning, Vol. 30, No. 2/3, pp. 271-274.

Kotsaintis, S., Kanellopoulos, D. 2006. Discretization Techniques: A recent survey. GESTS International Transactions on Computer Science and Engineering. Vol. 32(1), pp. 47-58.

Kurgan, L. A., and Cios, K. J. 2004. CAIM Discretization Algorithm. IEEE Transactions on Knowledge and Data Engineering. Vol. 16, No.2, pp. 145-152.

Kurgan, L. A., and Cios, K. J. 2003. Fast Class-Attribute Interdependence Maximization (CAIM) Discretization Algorithm. Proceeding of International Conference on Machine Learning and Applications. pp. 30-36.

Lan, W. H., and Frank, E. 2005. WEKA ( Waikato environment for knowledge analysis) . http://www.cs.waikato.ac.nz/ml/weka/ (accessed 17/03/13).

Liu, H., and Setiono, R. 1995. Chi2: Feature Selection and Discretization of Numeric Attribute. Proceeding of the IEEE 7th International Conference on Tools with Artificial Intelligence. pp.388-391.

Ming, X., and Xinping, X. 2009. A Comparative Analysis of Discretization Algorithms for Data Mining. Proceeding of 2009 IEEE International Conference on Grey Systems and Intelligent Services. November 10-12. pp. 1434-1438.

Monti, S., and Cooper, G. F. 1998. A multivariate discretization mrthod for learning bayesian networks from mixed data. In Proceedings on Uncertainty in Artificial Inlligence (UAI). pp. 404-413.

Peng, L., Qing, W., and Yujia, G. 2009. Study on Comparison of Discretization Methods. International Conference on Artificial Intelligence and Computational Intelligence. pp. 308-384.

Peng, L., Qing, W., and Yujia, G. 2009. Study on Comparison of Discretization Method. IEEE International Conference on Artificial Intelligence and Computational Intelligence. pp. 380-384.

Quinlan, J. R. 1993. C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc.

Radovanovic, M. 2006. Machine Learning in Web Mining. Master's thesis , Department of Mathematics and Informatics, Faculty of Science, University of Novi Sad, Serbia.

Roiger, J. R., and Geatz, M. W. 2003. Data Mining A Tutorial-Based Primer. Addison-Wesley Longman: Boston.

Sang, Y., Li, K., and Shen, Y. 2010. EBDA: An Effective Bottom-up Discretization Algorithm for Continuous Attributes. 10th IEEE International Conference on Computer and Information Technology (CIT 2010). pp. 2455-2462.

Senthilkumar, J., Manjula, D., and Krishnamoorthy, R. 2009. NANO: A New Supervised Algorithm for Feature Selection with Discretization. IEEE International Advance Computing Conference (IACC 2009), Patiala, India, pp. 1515-1520, 6-7 March.

Shiva, N., and Khare, M. 2004, Artificial neural network based line source models for vehicular exhaust emission predictions of an urban roadway. Transportation Research Part D: Transport and Environment . 9(May): 199-208.

Singh, P., and Verma, S. 2009. An Investigation of the Effect of Discretization on Defect Prediction Using Static Measure. IEEE International Conference on Advances in Computing, Control and Telecommunication Technologies. pp. 837-839.

Tsai, C. J., Lee, C. I., and Yang, W. P. 2008. A Discretization Algorithm Based on Class-Attribute Conticngency Coefficient. Information Sciences 178. pp. 174-731.

Wettayaprasit, W., and Nanakorn, P. 2006. Feature Extraction and interval Filtering Technique for Time -series Forecasting Using Neural Networke. IEEE Conferences Cybernetics and Intelligent Systems. pp. 635-640.

Wong A. K. C., and Chiu, D. K. Y. 1987. Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 9, pp. 796-805.

# ภาคผนวก

# ภาคผนวก ก

## ผลงานตีพิมพ์

| | |
|---|---|
| เรื่อง | การสร้างแบบจำลองพยากรณ์นาท่วมโดยใช้เทคนิคการทาเหมืองข้อมูลของอำเภอหาดใหญ่ |
| Conference | 2012 Ninth International Joint Conference on Computer Science and Software Engineering (JCSSE 2012) |
| สถานที่ | มหาวิทยลัยหอการค้า กรุงเทพมหานคร ประเทศไทย |
| วันที่ | 30 พค. – 1 มิย. 2555 |

# การสร้างแบบจำลองพยากรณ์น้ำท่วมโดยใช้เทคนิคการทำเหมืองข้อมูลของอำเภอหาดใหญ่

# Hatyai Flood Forecasting Model Using Data Mining Technique


**อับดุลเลาะ บากา [1] วิภาดา เวทย์ประสิทธิ์ [2] และศิริรัตน์ วณิชโยบล[3]**

**[1,2] ห้องปฏิบัติการปัญญาประดิษฐ์ [3]ห้องปฏิบัติการเทคโนโลยีระบบสารสนเทศและการวิจัยประยุกต์**

**ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ 15 ถ.กาญจนวณิชย์ อ.หาดใหญ่ จ.สงขลา 90110**

**[1]loh.bungraya@gmail.com , [2]wiphada.w@psu.ac.th , [3]sirirut.v@psu.ac.th**



## บทคัดย่อ

งานวิจัยนี้นำเสนอแบบจำลองการพยากรณ์น้ำท่วมโดยใช้เทคนิคการทำเหมืองข้อมูลด้วยต้นไม้ตัดสินใจ J48 โครงข่ายประสาทเทียม RBF และ MLP กรณีศึกษาอำเภอหาดใหญ่ จังหวัดสงขลา โดยใช้ข้อมูลน้ำท่วมของกรมชลประทานที่ 16 จังหวัดสงขลา จากสถานีโทรมาตรบ้านม่วงก็อง บ้านบางศาลา และบ้านหาดใหญ่ในของลุ่มน้ำคลองอู่ตะเภา อำเภอหาดใหญ่ ในการพยากรณ์ประเภทของการเตือนภัยน้ำท่วมในเขตเทศบาลอำเภอหาดใหญ่ (น้ำท่วม ธงแดง ธงเหลือง และธงเขียว) ผลลัพธ์ที่ได้แสดงให้เห็นว่าการใช้แบบจำลองการพยากรณ์น้ำท่วมโดยใช้เทคนิคการทำเหมืองข้อมูลด้วยต้นไม้ตัดสินใจ J48 ร่วมกับวิธีการแทนค่าข้อมูลที่สูญหายให้ค่าความถูกต้องสูงที่สุด นอกจากนี้ต้นไม้ตัดสินใจ J48 ยังแสดงความสัมพันธ์ของระดับความสูงของน้ำกับปริมาณการไหลของน้ำท่าในแต่ละสถานีโทรมาตรของลุ่มน้ำคลองอู่ตะเภาที่มีผลต่อการเกิดน้ำท่วม ซึ่งสามารถนำไปใช้ประโยชน์ในการบริหารจัดการน้ำได้อย่างมีประสิทธิภาพ



คำสำคัญ: การทำเหมืองข้อมูล, เครือข่ายประสาทเทียม, ต้นไม้ตัดสินใจ การพยากรณ์น้ำท่วม



## Abstract

This paper presents flood forecasting model by using data mining techniques of J48 decision tree RBF and MLP neural networks as a case study of Hat Yai, Songkla Provice. The data set of flood forecasting used for this study received from the three water stations of Ban Muangkong, Ban Bangsala, and Ban Hat Yai Nai of Klong U-Tapao basin under the 16th Irrigation Department in Songkla province. The type of flood forecasting are flood, red flag, yellow flag and green flag. The result of this study indicated that the pre-processing process prior entering to the data mining techniques by replacing missing values of J48 give highest accuracy value. Furthermore, J48 decision tree was able to show the rule of water quantity of each water station that would affect to flooding occurrence. Then will be new knowledge for efficient water management.



Keyword: Data Mining, Artificial Neural Networks, Decision Tree, Flood Prediction.


## 1. บทนำ

การเปลี่ยนแปลงภูมิอากาศของโลกอย่างรวดเร็วเป็นปัญหาหนึ่งที่สำคัญในสังคมปัจจุบัน โดยเฉพาะปัญหาทางธรรมชาติ เช่นอุทกภัย นับวันจะเพิ่มความรุนแรงมากขึ้นเรื่อยๆ การพยากรณ์น้ำท่วมและการค้นหาองค์ความรู้ของปัจจัยที่ทำให้เกิดน้ำท่วม เช่นการควบคุมปริมาณน้ำและทิศทางการไหลของน้ำ ดังนั้นการพยากรณ์น้ำท่วมจึงเป็นบทบาทหนึ่งที่ความสำคัญ อย่างไรก็ตามการพยากรณ์น้ำท่วมเป็นสิ่งที่ยุ่งยากซับซ้อน [1] เนื่องจากความสัมพันธ์ของกระบวนการทางอุทกวิทยากับข้อมูลน้ำท่วม มีลักษณะข้อมูลแบบไม่ใช่เชิงเส้น [2] อำเภอหาดใหญ่เป็นหนึ่งในเมืองเศรษฐกิจที่ใหญ่เป็นอันดับต้นๆ ของภาคใต้ แต่ด้วยลักษณะของพื้นที่เป็นแอ่งกระทะ และมีลุ่มน้ำคลองอู่ตะเภาไหลผ่านตัวเมืองหาดใหญ่ ลุ่มน้ำคลองอู่ตะเภาเป็นคลองหลักในการผันน้ำจากเขื่อนในอำเภอสะเดาลงสู่ทะเลสาบสงขลา จึงมีความเสี่ยงที่จะเกิดน้ำท่วมได้ง่ายและมีโอกาสทำความเสียหายต่อเศรษฐกิจนับพันล้านบาท จึงจำเป็นต้องมีการพัฒนาแบบจำลองที่ใช้ในการป้องกันน้ำท่วมที่มีประสิทธิภาพ ดังนั้นการนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้กับการสร้างแบบจำลองการพยากรณ์น้ำท่วมเป็นสิ่งที่จำเป็น เพื่อหาความสัมพันธ์ของปัจจัยต่างๆ ที่มีผลต่อการเกิดน้ำท่วม ทำให้สามารถทราบถึงเหตุการณ์ล่วงหน้าและมีการเตือนภัยได้อย่างมีประสิทธิภาพ [3] ผลลัพธ์ที่ได้แสดงความสัมพันธ์ของระดับความสูงของน้ำกับปริมาณการไหลของน้ำท่าในแต่ละสถานีโทรมาตรของลุ่มน้ำคลองอู่ตะเภา

บทความนี้นำเสนอแบบจำลองการพยากรณ์น้ำท่วมโดยใช้เทคนิคการทำเหมืองข้อมูล กรณีศึกษาอำเภอหาดใหญ่ จังหวัดสงขลา ตั้งแต่ปี พ.ศ. 2547-2554 ส่วนที่สองกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง ส่วนที่สามกล่าวถึงวิธีการดำเนินการวิจัย ส่วนที่สี่กล่าวถึงผลการทดลอง และสุดท้ายคือบทสรุป

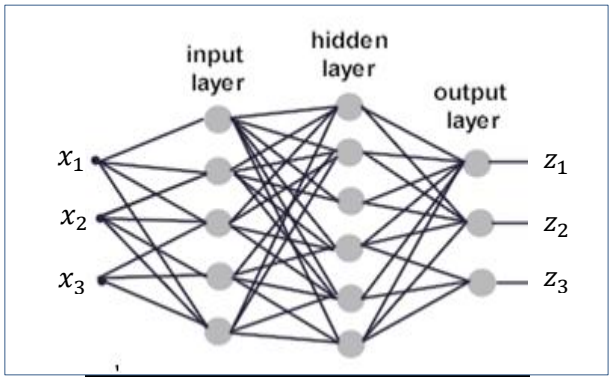
## 2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

### 2.1 J48 Algorithm

ต้นไม้ตัดสินใจเป็นวิธีการหนึ่งของการทำเหมืองข้อมูลที่ได้รับความนิยม เพราะง่ายต่อการทำความเข้าใจและมีความรวดเร็วในการประมวลผล มีลักษณะเหมือนโครงสร้างต้นไม้ แต่ละโหนดแสดงคุณลักษณะที่ใช้ในการทดสอบ แต่ละกิ่งแสดงผลในการทดสอบและโหนดใบแสดงกลุ่มหรือว่าคลาสที่กำหนดไว้ นอกจากนี้โครงสร้างของต้นไม้สามารถสร้างกฎเพื่อใช้ในการตัดสินใจได้ง่าย [4] และสามารถประมวลผลข้อมูลที่มีลักษณะเป็นตัวเลข (Numeric Data) ได้โดยไม่จำเป็นต้องมีการแบ่งช่วงข้อมูล (Discretization)

### 2.2 Artificial Neural Networks

โครงข่ายประสาทเทียม Artificial Neural Networks (ANN) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ ซึ่งมีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ [5] การเรียนรู้แบบมีผู้สอน (Supervise Learning) เพื่อให้คอมพิวเตอร์เรียนรู้รูปแบบและจดจำสิ่งที่เกิดขึ้นในอดีตเพื่อทำนายสิ่งที่จะเกิดขึ้นในอนาคต โครงข่ายประสาทเทียมประกอบด้วยหน่วยประมวลผลย่อยหรือเพอร์เซฟตรอน (Perceptron) หลายหน่วยเชื่อมต่อกันเป็นโครงข่าย โครงข่ายประสาทเทียมสามารถทำนายค่าที่มีความแม่นยำสูง ทนทานต่อความผิดพลาด และสามารถรองรับข้อมูลที่ไม่สมบูรณ์หรือมีสิ่งรบกวนได้ [6] นอกจากนี้ยังสามารถทำงานได้ดีกับข้อมูลแบบไม่ใช่เชิงเส้น โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นแสดงดังภาพที่ 1 ประกอบด้วยโครงข่าย 3 ชั้น คือ ชั้นข้อมูลเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นผลลัพธ์ (Output Layer) แต่ละหน่วยประมวลผลย่อยจะมีการคำนวณฟังก์ชันผลรวม (Summation Function) ซึ่งทำหน้าที่คำนวณหาผลรวมของผลคูณระหว่างค่าน้ำหนักกับค่าข้อมูลเข้าแสดงดังสมการที่ (1) ฟังก์ชันกระตุ้น (Activation Function) จะทำหน้าที่แปลงผลลัพธ์จากฟังก์ชันผลรวมให้อยู่ในช่วงที่ผู้ใช้ต้องการ




**ภาพที่ 1**: โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น

กำหนดให้ y คือ ฟังก์ชันผลรวม
$x_i$ คือ ค่าข้อมูลเข้าของนิวรอนตัวที่ i
$w_i$ คือ ค่าน้ำหนักของนิวรอนตัวที่ i
n คือ จำนวนนิวรอนของชั้นข้อมูลเข้า
b คือ ค่าความโน้มเอียง
z คือ ฟังก์ชันกระตุ้น

$$y = \sum_{i=0}^{n} x_i \; w_i + b \qquad (1)$$

$$z = \frac{1}{1+e^{-y}} \qquad (2)$$

$$z = e^{-y^2} \qquad (3)$$

โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (MLP) ใช้ฟังก์ชันกระตุ้นซิกมอยด์ (Sigmoid Function) แสดงดังสมการที่ (2) ส่วนโครงข่ายประสาทเทียมแบบ Radial Basis Function (RBF) ใช้ฟังก์ชันกระตุ้นเกาส์เซียน (Gaussian Function) แสดงดังสมการที่ (3)

### 2.3 งานวิจัยที่เกี่ยวข้อง

ปัจจุบันได้มีการพัฒนางานวิจัยด้านการทำเหมืองข้อมูลเพื่อสร้างแบบจำลองการพยากรณ์น้ำท่วมโดยใช้วิธีการต่างๆ McCulloch และคณะ [7] นำเสนอต้นไม้ตัดสินใจเพื่อใช้ในการพยากรณ์เวลาที่จะเกิดน้ำท่วมของประเทศอังกฤษ ซึ่งแบบจำลองที่ได้จากต้นไม้ตัดสินใจสามารถอธิบายความสัมพันธ์ของข้อมูลได้อย่างมีประสิทธิภาพ Cang และคณะ [8] ประยุกต์ใช้ทฤษฎี chaos ในการอธิบายพฤติกรรมของข้อมูลที่มีการเปลี่ยนแปลงตามกาลเวลาที่เปลี่ยนไป ทำให้ได้ผลการทดลองที่มีความถูกต้องสูงขึ้น และใช้ RBF ในการพยากรณ์อนุกรมเวลาความอลวน (Chaotic Time Series) เพื่อหาความสัมพันธ์ระหว่างจุดกำเนิดของการพยากรณ์และเวลาในการพยากรณ์ Xinhuan และ Zhuying [3] ประยุกต์ใช้โครงข่ายประสาทเทียมในการสร้างแบบจำลองการพยากรณ์น้ำท่วม เพื่อใช้ในการป้องกันและเตือนภัยน้ำท่วม แบบจำลองที่ได้ช่วยในการตัดสินใจได้อย่างมีประสิทธิภาพ Khalili และคณะ [9] นำเสนอแบบจำลองการ พยากรณ์ฝนตกรายวันของสถานี Mashhad ประเทศอิหร่านโดยใช้โครงข่ายประสาทเทียม เพื่อสกัดคุณลักษณะของข้อมูลน้ำฝน และสอนข้อมูลโดยใช้ Gradient descent algorithm ผลการทดลองแสดงให้เห็นว่า MLP มีประสิทธิภาพในการทำนายข้อมูลฝนตกรายวัน Wettayaprasit และ Nanakorn [10] ได้นำเสนอแบบจำลองการสกัดคุณลักษณะและเทคนิคการกรองข้อมูลสำหรับข้อมูลอนุกรมเวลาโดยใช้ MLP ในการทำนายค่าถัดไปที่เวลา t+1 ซึ่งผลลัพธ์แสดงให้เห็นว่า MLP มีประสิทธิภาพในการพยากรณ์อากาศได้ดี

## 3. วิธีการดำเนินการวิจัย

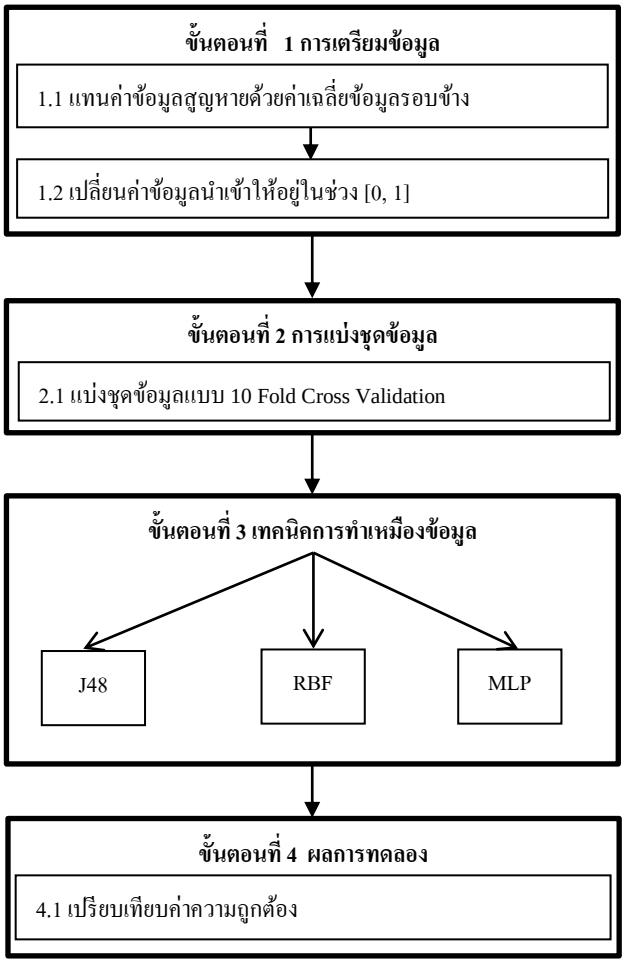
ขั้นตอนในการสร้างแบบจำลองการพยากรณ์น้ำท่วมโดยใช้เทคนิคของการทำเหมืองข้อมูลมีทั้งหมด 4 ขั้นตอนแสดงดังภาพที่ 2 โดยมีรายละเอียดดังต่อไปนี้

**ขั้นตอนที่ 1** การเตรียมข้อมูล (Data Preprocessing) คือ การแทนค่าข้อมูลที่สูญหาย (Replace Missing Values) ด้วยค่าเฉลี่ยของข้อมูลรอบข้างซึ่งสามารถคำนวณได้ดังสมการที่ (4) และทำการเปลี่ยนค่าข้อมูลเข้า (Data Transformation) ให้อยู่ในช่วงค่า [0, 1] แสดงดังสมการที่ (5)

**ขั้นตอนที่ 2** การแบ่งชุดข้อมูลในการสร้างแบบจำลอง ใช้วิธีการแบ่งชุดข้อมูลออกเป็น 10 ชุด ใช้ข้อมูล 9 ชุดในการสอนและข้อมูล 1 ชุดในการทดสอบ โดยชุดข้อมูลแต่ละชุดจะถูกสอนและทดสอบสลับกันไป เรียกวิธีการนี้ว่า 10-Fold Cross Validation

**ขั้นตอนที่ 3** เทคนิคที่ใช้ในการทำเหมืองข้อมูลคือ ต้นไม้ตัดสินใจ J48 โครงข่ายประสาทเทียม RBF และ MLP ในการสร้างแบบจำลองการพยากรณ์น้ำท่วม

**ขั้นตอนที่ 4** การเปรียบเทียบผลการทดลอง โดยใช้วิธีการประเมินค่าความถูกต้อง (Accuracy) ในการพยากรณ์ดังแสดงในสมการที่ (6)




**ภาพที่ 2:** ขั้นตอนการสร้างแบบจำลองพยากรณ์น้ำท่วม

$$\text{ค่าข้อมูลสูญหาย} = \frac{\text{ค่าก่อนหน้าข้อมูลหาย+ค่าหลังข้อมูลหาย}}{2} \quad (4)$$

$$\text{ข้อมูลใหม่ในการเปลี่ยนข้อมูล} = \frac{\text{ค่าข้อมูลเดิม-ค่าต่ำสุดในช่วง}}{\text{ค่าสูงสุดในช่วง-ค่าต่ำสุดในช่วง}} \quad (5)$$

$$\text{ค่าความถูกต้อง} = \frac{\text{จำนวนข้อมูลที่ทำนายถูก}}{\text{จำนวนข้อมูลทั้งหมดในคลาส}} \times 100 \quad (6)$$

ชุดรูปแบบที่ใช้ในการทดลองได้แบ่งออกเป็น 4 ชุดการทดลองคือ ชุดการทดลอง A เป็นชุดการทดลองที่ใช้ข้อมูลดิบ (ข้อมูลที่มี Missing Values) ชุดการทดลอง B เป็นชุดการทดลองที่ใช้ข้อมูลที่มีการแทนค่าข้อมูลที่สูญหายเพียงอย่างเดียว ชุดการทดลอง C เป็นชุดการทดลองที่ใช้ข้อมูลดิบจากแบบจำลอง A แล้วเปลี่ยนค่าข้อมูลให้อยู่ในช่วง [0, 1] และชุดการทดลอง D เป็นชุดการทดลองที่ใช้ข้อมูลจากชุดการทดลอง B แล้วเปลี่ยนค่าข้อมูลให้อยู่ในช่วง [0, 1] แสดงดังตารางที่ 1 โดยทั้ง 4 ชุดการทดลองจะใช้เทคนิคการทำเหมืองข้อมูล 3 แบบคือต้นไม้ตัดสินใจ J48 โครงข่ายประสาทเทียม RBF และ MLP โดยใช้ซอฟต์แวร์ WEKA Version 3.6 ในการทดลอง

**ตารางที่ 1:** รูปแบบการทดลอง

| ชุดการทดลอง | ReplaceMissing Value | Data Transfor-mation | J48 | RBF | MLP |
|---|---|---|---|---|---|
| A | × | × | ✓ | ✓ | ✓ |
| B | ✓ | × | ✓ | ✓ | ✓ |
| C | × | ✓ | ✓ | ✓ | ✓ |
| D | ✓ | ✓ | ✓ | ✓ | ✓ |

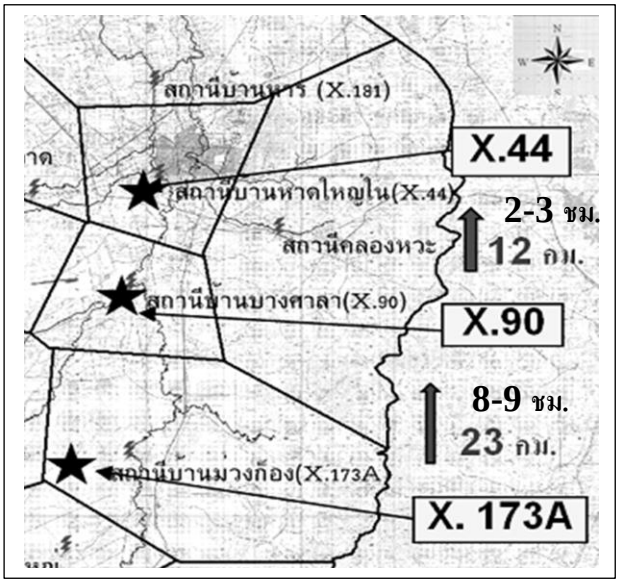
## 4. ผลการทดลอง

ข้อมูลที่ใช้ในงานศึกษาวิจัยครั้งนี้ประกอบด้วยชุดข้อมูล 2 ชุดคือ ชุดที่ 1 คือข้อมูลน้ำท่วมรายชั่วโมงและรายวัน อำเภอหาดใหญ่ เป็นข้อมูลที่ได้จากกรมชลประทานที่ 16 จังหวัดสงขลา และชุดที่ 2 ข้อมูลฝนตกรายวันอำเภอหาดใหญ่ จังหวัดสงขลา เป็นข้อมูลที่ได้จากกรมอุตุนิยมวิทยาภาคใต้ฝั่งตะวันออก จังหวัดสงขลา โดยมีรายละเอียดดังนี้

4.1 ข้อมูลน้ำท่วมอำเภอหาดใหญ่ตั้งแต่ปี พ.ศ. 2547-2554 โดยเก็บข้อมูลปริมาณน้ำฝน ปริมาณการไหลของน้ำท่า และความสูงของระดับน้ำจากสถานีโทรมาตร 3 สถานีของคลองอู่ตะเภา คือสถานีบ้านม่วงก็อง สถานีบ้านบางศาลา และสถานีบ้านหาดใหญ่ใน โดยทิศทางการเดินทางของน้ำเริ่มต้นจากสถานีบ้านม่วงก็อง (X.173A) ไปยังสถานีบ้านบาง

**ตารางที่ 2:** ตัวอย่างข้อมูลน้ำท่วม อำเภอหาดใหญ่

| ปริมาณน้ำฝนในแต่ละสถานี (RF) (มิลลิเมตร) | | | ปริมาณการไหลของน้ำท่าในแต่ละสถานี (D) (ลูกบาศก์เมตร/วินาที) | | | ระดับความสูงของน้ำในแต่ละสถานี (WL) (เมตร) | | ประเภทของการเตือนภัย |
|---|---|---|---|---|---|---|---|---|
| ม่วงก็อง | บางศาลา | หาดใหญ่ใน | ม่วงก็อง | บางศาลา | หาดใหญ่ใน | ม่วงก็อง | บางศาลา | |
| RF-X.173A | RF-X.90 | RF-X.44 | D-X.173A | D-X.90 | D-X.44 | WL-X.173A | WL-X.90 | Class |
| 22 | 2 | 18 | 430.2 | 738.3 | 572.0 | 16.53 | 8.47 | Flood |
| 12 | 0 | 1 | 124.8 | 266.7 | 272.2 | 16.51 | 7.28 | Red |
| 1 | 3 | 8 | 27.8 | 215.4 | 229.7 | 14.47 | 7.15 | Yellow |
| 0 | 2 | 4 | 33.55 | 70.2 | 95.8 | 13.59 | 4.14 | Green |



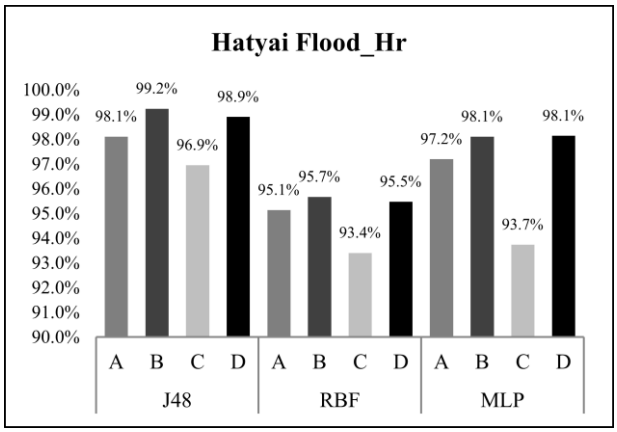

**ภาพที่ 3:** แผนที่สถานีโทรมาตรของอู่ตะเภา อำเภอหาดใหญ่ [11]

ศาลา (X.90) มีระยะทางตามลำน้ำ 23 กิโลเมตร ใช้เวลาในการเดินทางของน้ำประมาณ 8-9 ชั่วโมง และจากสถานีบ้านบางศาลา (X.90) ไปยังสถานีเป้าหมายคือสถานีบ้านหาดใหญ่ใน (X.44) มีระยะทางตามลำน้ำ 12 กิโลเมตร เวลาในการเดินทางของน้ำประมาณ 2-3 ชั่วโมงแสดงดังภาพที่ 3 [11]

ข้อมูลน้ำท่วมที่ใช้ในการทดลองแบ่งออกเป็นอีก 2 ชุดย่อยคือ 1) ชุดข้อมูลน้ำท่วมรายชั่วโมง มีจำนวนทั้งหมด 38,477 รายการ มีการสุ่มเลือกตัวอย่างมา 3,022 รายการ และ 2) ชุดข้อมูลน้ำท่วมรายวัน มีจำนวนทั้งหมด 2,430 รายการ ทั้งสองชุดข้อมูลประกอบด้วยข้อมูลเข้า 8 แอตทริบิวต์แสดงดังตารางที่ 2 คือข้อมูลปริมาณน้ำฝนสถานีบ้านม่วงก็อง (RF-X.173A) ปริมาณน้ำฝนสถานีบ้านบางศาลา (RF-X.90) ปริมาณน้ำฝนบ้านหาดใหญ่ใน (RF-X.44) ปริมาณการไหลของน้ำท่าสถานีบ้านม่วงก็อง (D-X.173A) ปริมาณการไหลของน้ำท่าสถานีบ้านบางศาลา (D-X.90) ปริมาณการไหลของน้ำท่าสถานีบ้านหาดใหญ่ใน (D-X.44)

**ตารางที่ 3:** ประเภทของการเตือนภัยน้ำท่วม อำเภอหาดใหญ่ [11]

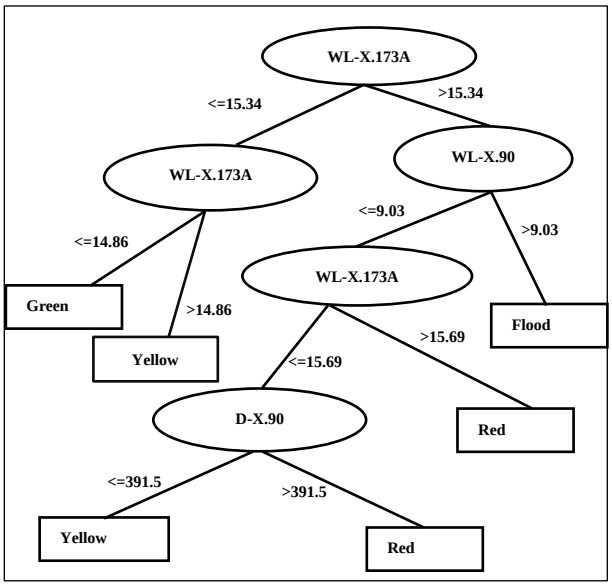
| ประเภทการเตือนภัย | ความสูงของระดับน้ำ (เมตร) | ความหมาย |
|---|---|---|
| Flood | มากกว่า 7.2 | น้ำท่วมภายใน 3 ชม |
| Red | อยู่ระหว่าง 5.7 – 7.2 | ธงแดงเตรียมอพยพ |
| Yellow | อยู่ระหว่าง 4.2 – 5.6 | ธงเหลืองเฝ้าระวัง |
| Green | น้อยกว่า 4.2 | ธงเขียวปลอดภัย |




**ภาพที่ 4:** ผลการทดสอบชุดข้อมูลน้ำท่วมรายชั่วโมง อำเภอหาดใหญ่

ระดับความสูงของน้ำสถานีบ้านม่วงก็อง (WL-X.173A) และระดับความสูงของน้ำสถานีบ้านบางศาลา (WL-X.90) และใช้ความสูงของระดับน้ำจากสถานีบ้านหาดใหญ่ในเป็น Target Class เนื่องจากเป็นสถานีที่ใช้ในการ แจ้งเตือนภัยน้ำท่วมในปัจจุบัน โดยมีผู้เชี่ยวชาญกำหนดประเภทของการเตือนภัยน้ำท่วม 4 แบบคือ น้ำท่วม (Flood) ธงแดง (Red) ธงเหลือง (Yellow) และธงเขียว (Green) แสดงดังตารางที่ 3

ผลการทดลองชุดข้อมูลน้ำท่วมรายชั่วโมง อำเภอหาดใหญ่ สรุปได้ว่าแบบจำลองที่มีการแทนค่าข้อมูลที่สูญหายให้ค่าความถูกต้องสูงกว่าแบบจำลองที่ใช้ข้อมูลดิบ กล่าวคือ J48 แบบจำลอง B ให้ค่าความถูกต้องสูงสุด 99.2% RBF แบบจำลอง B ได้ค่าความถูกต้องสูงสุดคือ 95.7% ส่วน MLP ทั้งแบบจำลอง B และ D ให้ค่าความถูกต้องสูงสุดคือ




ภาพที่ 5: ต้นไม้ตัดสินใจข้อมูลน้ำท่วมรายชั่วโมง อำเภอหาดใหญ่

**R1:** ถ้าระดับน้ำจากสถานีบ้านม่วงก็องมากกว่า 15.34 เมตร และระดับน้ำจากสถานีบ้านบางศาลามากกว่า 9.03 เมตร แล้วทำนายว่า **น้ำท่วม**

**R2:** ถ้าระดับน้ำจากสถานีบ้านม่วงก็องมากกว่า 15.69 เมตร และระดับน้ำจากสถานีบ้านบางศาลาน้อยกว่าหรือเท่ากับ 9.0 เมตร แล้วทำนายว่า **ธงแดง**

**R3:** ถ้าระดับน้ำจากสถานีบ้านม่วงก็องอยู่ระหว่าง 15.34 เมตร ถึง 15.69 เมตรและระดับน้ำจากสถานีบ้านบางศาลาน้อยกว่าหรือเท่ากับ 9.03 เมตร และมีปริมาณการไหลของน้ำท่ามากกว่า 391.5 ลูกบาศก์เมตรต่อวินาที แล้วทำนายว่า **ธงแดง**

**R4:** ถ้าระดับน้ำจากสถานีบ้านม่วงก็องอยู่ระหว่าง 15.34 เมตร ถึง 15.69 เมตรและระดับน้ำจากสถานีบ้านบางศาลาน้อยกว่าหรือเท่ากับ 9.03 เมตร และมีปริมาณการไหลของน้ำท่าน้อยกว่าหรือเท่ากับ 391.5 ลูกบาศก์เมตรต่อวินาที แล้วทำนายว่า **ธงเหลือง**

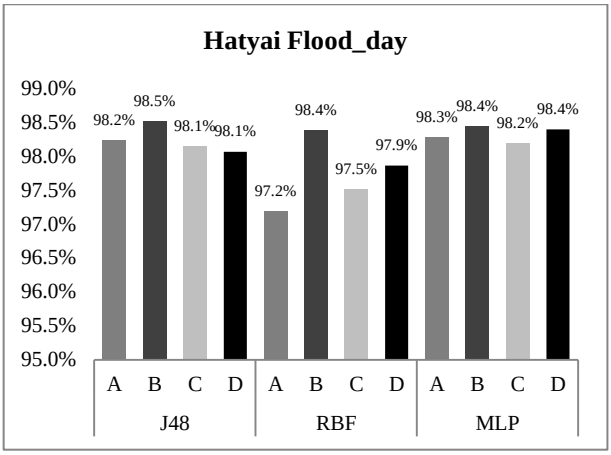
**R5:** ถ้าระดับน้ำจากสถานีบ้านม่วงก็องอยู่ระหว่าง 14.86 เมตร ถึง 15.34 เมตร แล้วทำนายว่า **ธงเหลือง**

**R6:** ถ้าระดับน้ำจากสถานีบ้านม่วงก็องน้อยกว่าหรือเท่ากับ 14.86 เมตร แล้วทำนายว่า **ธงเขียว**

ภาพที่ 6: กฎความสัมพันธ์ของข้อมูลน้ำท่วม อำเภอหาดใหญ่

98.1% แสดงดังภาพ ที่ 4 ในขณะที่แบบจำลอง C ที่มีการเปลี่ยนค่าข้อมูลให้อยู่ในช่วง [0, 1] เพียงอย่างเดียวของ J48 RBF และ MLP ได้ค่าความถูกต้องต่ำสุด

ต้นไม้ตัดสินใจ J48 ของข้อมูลน้ำท่วมรายชั่วโมงแสดงได้ดังภาพที่ 5 และสามารถแปลงเป็นกฎความสัมพันธ์ได้ 6 กฎแสดงดังภาพที่ 6




ภาพที่ 7: ผลการทดลองข้อมูลน้ำท่วมรายวัน อำเภอหาดใหญ่

ตัวอย่างเช่น กฎ R1 ถ้าความสูงของระดับน้ำที่สถานีบ้านม่วงก็องมากกว่า 15.34 เมตร และความสูงของระดับน้ำที่สถานีบ้านบางศาลามากกว่า 9.03 เมตร แล้วทำนายว่าน้ำท่วม เป็นต้น

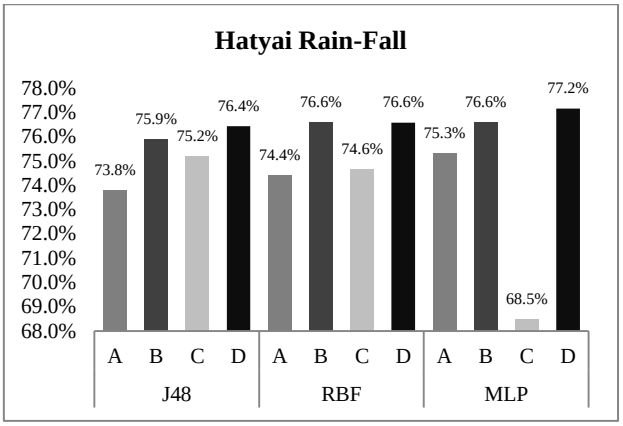
ผลการทดลองข้อมูลน้ำท่วมรายวันของอำเภอหาดใหญ่ J48 แบบจำลอง B ได้ค่าความถูกต้องสูงสุดคือ 98.5% RBF และ MLP แบบจำลอง B ได้ค่าความถูกต้องสูงสุดคือ 98.4% แบบจำลอง D ของ MLP ได้ค่าความถูกต้องสูงสุด 98.4% เช่นกัน ในขณะที่แบบจำลอง C ที่มีการเปลี่ยนค่าข้อมูลให้อยู่ในช่วง [0, 1] อย่างเดียวของ J48 RBF และ MLP ได้ค่าต่ำสุดแสดงดังภาพที่ 7

สรุปผลการทดลองชุดข้อมูลน้ำท่วมรายชั่วโมงและน้ำท่วมรายวันอำเภอหาดใหญ่แสดงให้เห็นว่า แบบจำลอง B ที่มีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้างจะได้ค่าความถูกต้องสูงที่สุดเมื่อเทียบกับแบบจำลองอื่นๆ โดยเฉพาะเมื่อใช้ร่วมกับต้นไม้ตัดสินใจ J48 จะได้ค่าความถูกต้องสูงที่สุด รองลงมาเป็น MLP และ RBF ตามลำดับ

4.2 ข้อมูลฝนตกรายวันรายวันประกอบด้วยข้อมูลเข้า 3 แอตทริบิวต์คือ ความเร็วลมสูงสุด (Maxwind) ความชื้นสัมพัทธ์ (RH) และอุณหภูมิ (Temp) และใช้ปริมาณน้ำฝนเป็น Target Class โดยมีการแบ่งประเภทข้อมูลออกเป็นสองกลุ่มคือ ถ้าปริมาณน้ำฝนมากกว่า 0 หมายถึงฝนตก ถ้าปริมาณน้ำฝนเท่ากับ 0 หมายถึงฝนไม่ตก ตัวอย่างข้อมูลแสดงดังตารางที่ 4

**ตารางที่ 4:** ตัวอย่างข้อมูลฝนตกรายวัน อำเภอหาดใหญ่

| Maxwind (km/h) | RH (%) | Temp (°C) | Class |
|---|---|---|---|
| 7 | 85 | 27.1 | ฝนตก |
| 6 | 93 | 25.4 | ฝนตก |
| 10 | 85 | 27.0 | ฝนไม่ตก |
| 9 | 82 | 27.3 | ฝนไม่ตก |




ภาพที่ 8: ผลการทดลองข้อมูลฝนตกรายวัน อำเภอหาดใหญ่

ผลการทดลองข้อมูลฝนตกรายวัน แสดงให้เห็นว่า MLP ได้ค่าความถูกต้องสูงสุดคือ แบบจำลอง D คือ 77.2% RBF ได้ค่าความถูกต้องสูงสุดคือแบบจำลอง B และแบบจำลอง D คือ 76.6% ส่วน J48 แบบจำลอง D ได้ค่าความถูกต้องสูงสุดคือ 76.4% แสดงดังภาพที่ 8 และสำหรับชุดข้อมูลที่ได้ค่าความถูกต้องต่ำสุดของ J48 และ RBF คือแบบจำลอง A ที่ใช้ข้อมูลดิบคือ 73.8% และ 74.4 % ตามลำดับ ส่วน MLP ได้ค่าความถูกต้องต่ำสุดจากแบบจำลอง C คือ 68.5%

## 5. สรุป

จากข้อมูลน้ำท่วมรายชั่วโมงและรายวัน อำเภอหาดใหญ่ จังหวัดสงขลา สามารถสรุปได้ดังนี้ 1) ชุดข้อมูลที่มีการแทนค่าข้อมูลที่สูญหายด้วยค่าเฉลี่ยรอบข้าง (แบบจำลอง B) มีผลทำให้ค่าความถูกต้องสูงกว่าแบบจำลองอื่นๆ โดยเฉพาะเมื่อมีการจำแนกข้อมูลโดยใช้ต้นไม้ตัดสินใจ J48 จะมีค่าความถูกสูงกว่าเทคนิคอื่นๆ 2) ชุดข้อมูลที่มีการแทนค่าข้อมูลที่สูญหายและการเปลี่ยนค่าข้อมูลให้อยู่ช่วง [0, 1] (แบบจำลอง D) มีผลให้ค่าความถูกต้องสูงกว่าเมื่อเทียบกับชุดข้อมูลดิบ 3) ชุดข้อมูลที่มีการเปลี่ยนค่าข้อมูลให้อยู่ในช่วง [0,1] แต่ไม่มีการแทนค่าที่สูญหาย (แบบจำลอง C) จะให้ค่าความถูกต้องน้อยลงเมื่อเทียบกับข้อมูลดิบ ดังนั้นจึงไม่ควรที่จะทำการเปลี่ยนค่าข้อมูลให้อยู่ในช่วง [0, 1] ในข้อมูลดิบที่มีค่าที่สูญหาย นอกจากนี้ต้นไม้ตัดสินใจ J48 สามารถแสดงความสัมพันธ์ของระดับความสูงของน้ำกับปริมาณการไหลของน้ำท่าในแต่ละสถานีโทรมาตรของคลองอู่ตะเภา อำเภอหาดใหญ่ ที่มีผลต่อการเกิดน้ำท่วมล่วงหน้า ทำให้มีเวลาในการเตรียมความพร้อมรับมือกับภัยพิบัติที่จะเกิดขึ้น อีกทั้งยังช่วยให้ผู้ที่เกี่ยวข้องในการบริหารจัดการน้ำสามารถจัดการน้ำได้อย่างมีประสิทธิภาพต่อไป

## เอกสารอ้างอิง


[1] C. Zhu and X. Ma, “Simulation of Flood Water Level Using PSO-based RBF Neural Network”, the Third International Symposium on Intelligence Information Technology Application, pp. 68-71, 2009.

[2] G. Corani and G. Guariso, “Coupling Fuzzy Modeling and Neural Networks for River Flood Prediction”, IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, vol. 35, no. 3, pp. 382-388, August 2005.

[3] C. Xinhua and L. Zhuying, “The Application of Neural Network Technology in Floodwater Forecast”, International Conference on Networking and Digital Society, pp. 419-421, 2010.

[4] A. Cufoglu, M. Lohi and K. Madani, “A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling”, World Congress on Computer Science and Information Engineering, pp. 708-712, 2009.

[5] R. J. Roiger and M. W. Geatz, Data mining a Tutorial-Based Primer, Pearson Education, Inc., 2003.

[6] เสาวลักษณ์ อร่ามพงศานุวัต และพยุง มีสัจ, “การพัฒนาแบบจำลองเพื่อพยากรณ์ปริมาณ PM10 ในเขตกรุงเทพมหานครโดยใช้โครงข่ายประสาทเทียม”, ประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 6 พ.ศ. 2553 หน้า 104-109.

[7] D. R. McCulloch, J. Lawry and I. D. Cluckie, “Real-Time Flood Forecasting Using Updateable Linguistic Decision Trees”, IEEE International Coference on Fuzzy System, pp. 1935-1942, 2008.

[8] X. J. Cang and et al, “A method of Flood Forecasting of Chaotic Radial Basis Function Neural Network”, the 2nd International Workshop on ISA, pp. 1-5, 2010.

[9] N. Khalili and et al, “Daily Rainfall Forecasting for Mashhad Synoptic Station Using Artificial Neural Networks”, International Conference on Environmental and Computer Science, vol. 19, pp. 118-123, 2011.

[10] W. Wettayaprasit and P. Nanakorn, “Feature Extraction and Interval Filtering Technique for Time-Series Forecasting Using Neural Networks”, IEEE Conferences on Cybernetics and Intelligent Systems, pp. 635-640, 2006.

[11] กรมชลประทานที่ 16 จังหวัดสงขลา. คู่มือการเฝ้าระวังและการเตือนภัยน้ำท่วมอำเภอหาดใหญ่. 2553.

# ภาคผนวก ข

## ผลงานตีพิมพ์

| | |
|---|---|
| เรื่อง | A Novel Discretization Technique Using Class Attribute Interval Average |
| Conference | The Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP2014) |
| สถานที่ | มหาวิทยลัยหอการค้า กรุงเทพมหานคร ประเทศไทย |
| วันที่ | 8-6 พค. 2557 |

# A Novel Discretization Technique Using Class Attribute Interval Average

Abdulloh Baka
Artificial Intelligence Research LAB,
Department of Computer Science,
Prince of Songkla University,
Hat Yai, Songkhla, Thailand
loh.bungaraya@gmail.com

Wiphada Wettayaprasit
Artificial Intelligence Research LAB,
Department of Computer Science,
Prince of Songkla University,
Hat Yai, Songkhla, Thailand
wiphada.w@psu.ac.th

Sirirut Vanichayobon
Information Systems Technology and
Applied Research LAB,
Department of Computer Science,
Princeof Songkla University,Thailand
sirirut.v@psu.ac.th

*Abstract* – **Discretization algorithm is important for data mining preprocessing because it will help the user to easily understand the data, reduce the complexity of data, reduce processing time, and increase efficiency and accuracy of the data. This paper proposes the new discretization algorithm called Class Attribute Interval Average (CAIA). The algorithm uses 2D-quanta matrix table to calculate each of class individual interval's average and merge the best adjacent intervals to form the new interval. The experimental design uses four-UCI data sets (Iris, Breast Cancer, Heart Diseases, Glass) and four-classification algorithms (J48, RBF, MLP, NB). The comparisons of experimental result with the other six discretization algorithms (EW, EF, ChiMerge, IEM, CAIM, CACC) show that the proposed CAIA has the best mean rank for both of the accuracy and the number of intervals.**

***Keywords- Data Discretization; Data Mining; Classification***

## I. INTRODUCTION

Data mining is the process of extracting the information from data warehouse or database, which refers to data in the past to predict the tendency and the action of data in the future [1]. Machine learning algorithms such as decision tree, neural networks, and naïve bay are also used to find the rules or pattern of data. Normally, data is divided into two types: 1) discrete data and 2) continuous data. Discrete data is a qualitative data such as gender and education level while continuous data is a quantitative data such as height and age. Most real world application usually involves with continuous data [3].

Data preprocessing is important for data mining because the quality of data mining result is usually depended on the quality of the data preprocessing. The data preprocessing steps are data discretization, data transformation, and data reduction [2]. Some data mining algorithms use only discrete data, so that the process to change the continuous data into discrete data is needed. This process is called discretization. The advantages of discretization are 1) higher accuracy 2) faster execution time 3) easy to understand data, and 4) change continuous data into discrete data [2, 4, 5]. The examples of discretization algorithms are Equal-Width (EW) [6] and Equal-Frequency (EF) [6], ChiMerge [7], Information Entropy Maximization (IEM) [12], Class-Attribute Interdependence Maximization (CAIM) [4] and Class-Attribute Contingency Coefficient (CACC) [2].

The problem of discretization algorithm, such as EW and EF, is that users need to specify the parameters for the stopping criterion. Although these processes are fast and easy, it is difficult to specify the suitable parameters for each given data set [6]. This is the same as ChiMerge algorithm when users have to specify level of significance for the stopping criterion [7]. CAIM considers only the class with the most samples and ignore all other target classes that will decrease the quality of discretization in some cases [2]. However, CACC will spend more time for calculation because the process is quite complicate. The experimental results in [1, 3, 4, 11] show that the number of intervals close to the number of classes can give the high accuracy.

Section II of this paper will mention about related work. Section III is the proposed Class Attribute Interval Average (CAIA) algorithm. Section IV is the experiment result. The last section is the conclusions.

## II. RELATED WORK

This section has presented about theory and research in term of discretization process. Since 1987, discretization method has been divided into 3 methods of 1) binning 2) statistical, and 3) information entropy. Binning method is the simplest method. User needs to specify the number of bins. The examples of this binning method are EW and EF [6]. Statistical method is a process using statistic theory to find the cutting point for splitting or merging of the intervals [10]. Information entropy method is a process that uses the entropy value for evaluating the disorder of data. If the data has high disorder, then the entropy value will be high. If the data has low disorder (high regularity of data), then the entropy value will be low. Therefore, the discretization process will use the lowest entropy value to find the cutting point for splitting or merging of the intervals [3].

ISBN: 978-1-4799-3723-3 

TABLE 1. REVIEWED OF DISCRETIZATION ALGORITHMS.

| Year | Algorithms | Discretization Methods | Characteristics | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Unsupervised | Supervised | Multivariate | Univariate | Global | Local | Dynamic | Static | Split | Merge |
| 1987 | Equal-width [6] | binning | ✓ | × | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| 1987 | Equal-frequency [6] | binning | ✓ | × | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| 1992 | ChiMerge [7] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| 1993 | IEM [12] | information entropy | × | ✓ | × | ✓ | × | ✓ | × | ✓ | ✓ | × |
| 1995 | Chi2 [8] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| 1998 | Multi-Bayesian [15] | information entropy | × | ✓ | ✓ | × | × | ✓ | ✓ | × | ✓ | × |
| 2004 | CAIM [4] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| 2003 | F-CAIM [13] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| 2006 | FastICA [16] | statistical | × | ✓ | ✓ | × | × | ✓ | × | ✓ | × | ✓ |
| 2008 | CACC [2] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | × |
| 2009 | SX-Mean [14] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| 2010 | EBDA [9] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |
| 2011 | NCL-CAIR [11] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | ✓ | ✓ |
| 2013 | CAIA [proposed] | statistical | × | ✓ | × | ✓ | ✓ | × | × | ✓ | × | ✓ |

The review of discretization algorithms can be divided into 5 characteristics as follows: 1) unsupervised versus supervised, 2) multivariate versus univariate, 3) global versus local, 4) dynamic versus static, and 5) splitting versus merging as shown in Table 1.

The explanations for the five characteristics are as follows. 1) Unsupervised learning does not need class for consideration while supervised learning does need class for discretization. Most discretization processes proposed in the literature are supervised and can be applied over supervised data mining problems. 2) Multivariate will consider more than one attributes in the discretization process to select the cutting point for splitting or for merging while univariate will proceed only one attribute to select the cutting point for splitting or for merging. 3) Global will use the total instances to generate in the discretization process while local will use only some parts of instances to generate in the discretization process. 4) Dynamic will do the discretization proceed at the same time as the learning method in data mining while static will do the discretization process before the learning method in data mining. 5) Splitting is a top-down process. It begins with one interval and continue selecting the new cutting point until the stopping condition is met. On the other hand, merging is a bottom-up process. It will identify the cutting point of all data and continue selecting the new merging point until the stopping condition is met.

From Table 1, the discretization methods using information entropy are IEM [12] and Multi-Bayesian [15]. IEM [12] is the supervised, univariate, global, static, and merge characteristics. Multi-Bayesian [15] is supervised, mutivariate, local, dynamic, and split characteristics. In 1987, Wong and Chiu presented unsupervised discretization algorithms Equal-Width (EW) [6] and Equal-Frequency (EF) [6]. Both algorithms are binning method. EW used the same size of range for discretization process. EF used the same frequency for discretization process. User has to select the parameter for a stopping criteria. Even though this process is a fast and easy process, it is difficult to find the suitable parameter for the data set. In 1992, Kerber presented supervised discretization algorithm called ChiMerge [7] which used $x^2$ statistic value to find the number of intervals for the discretization process. The $x^2$ threshold value was used to select to merge with the adjacent interval.

Most algorithms will be either splitting or merging process, but NCL-CAIR [11] will use both splitting and merging. For merging process in 1995, Liu and Setiono proposed Chi2 [8] to solve the ChiMerge [7] problem where user had to specify level of significance before discretization process. Chi2 [8] will use data inconsistency to perform the automatic stopping criteria. In 2010, Sang and et al. proposed EBDA [9] which developed from Chi2 [8] for better merging criteria. For splitting process, in 2004, Kurgan and Cios proposed CAIM [4] which used interdependence maximization criteria between class and attribute for splitting data. In 2008, Tsai and et al. proposed CACC [2] which considered of every instance that has been discretized to find the interdependence maximization value between class and attribute to find the best cutting point for discretization process.

## III. CLASS ATTRIBUTE INTERVAL AVERAGE ALGORITHM

Class Attribute Intervals Average (CAIA) is the new proposed discretization algorithm. The CAIA is divided into 2 main steps as shown in Fig. 1. The first step is to create a 2D-quanta matrix. The second step is to merge the intervals.

***Given:*** Data set where M is the total number of instances, s is the total number of classes, and x is the total number of attributes

***Step1 . Create 2D-quanta matrix***

1.1 For p = 1 to x (each continuous attribute $A_p$)
1.2 Let $d_1$ = minimum value of $A_p$
1.3 Let $d_n$ = maximum value of $A_p$
1.4 Sort all distinct values of $A_p$ in ascending order.
1.5 Merge all adjacent intervals that belongs to the same class.
1.6 For r = 1 to n // n is the number of intervals
1.7 //Calculate midpoint of each adjacent intervals
$B_{+r} = (Bmax_r + Bmin_{r+1})/2$
1.8 Output is the 2D-quanta matrix, discretization scheme D for attribute $A_p$.

***Step2 . Merge using Class Attribute Interval Average Algorithm***

2.1 For r = 1 to n // n is the total number of intervals
2.2 For i = 1 to s
2.3 // Calculate Class Attribute Interval
$CAI_{+r} = \sum_{i=1}^{s} (q_{ir}^2/M_{i+}) * (M_{+r})$
2.4 If (n > s) then
2.5 $CAI_{min} = CAI_{+r}$ minimum value
2.6 candidate merge $CAI_{min}$ with $CAI_{left\_of\_min}$,
2.7 $CAIA_{left} = \sum_{r=1}^{n} CAI_{+r}/n$
2.8 candidate merge $CAI_{min}$ with $CAI_{right\_of\_min}$,
2.9 $CAIA_{right} = \sum_{r=1}^{n} CAI_{+r}/n$
2.10 If ($CAIA_{left} > CAIA_{right}$)
2.11 then $CAIA = CAIA_{left}$
2.12 else $CAIA = CAIA_{right}$
2.13 n = n - 1
2.14 Output discretization scheme D' for attribute $A_p$.

Fig. 1. Class Attribute Intervals Average Algorithm.

Step 1 is to create a 2D-quanta matrix values. By explaining how the algorithm works, Fig. 2 will show the example of age attribute with the number of instances M is 30, the number of classes s is 3 (Care, Edu and Work), and the number of attributes x is 1.

TABLE 2. 2D-QUANTA MATRIX OF ATTRIBUTE $A_p$.

| Class | Interval | | | | | Sum of class |
|---|---|---|---|---|---|---|
| | 1 | … | r | … | n | |
| | $[d_0,d_1]$ | … | $(d_{r-1},d_r]$ | … | $(d_{n-1},d_n]$ | |
| $C_1$ | $q_{11}$ | … | $q_{1r}$ | … | $q_{1n}$ | $M_{1+}$ |
| $C_i$ | $q_{i1}$ | … | $q_{ir}$ | … | $q_{in}$ | $M_{i+}$ |
| $C_s$ | $q_{s1}$ | … | $q_{sr}$ | … | $q_{sn}$ | $M_{s+}$ |
| **Sum of interval** | $M_{+1}$ | … | $M_{+r}$ | … | $M_{+n}$ | M |
| $CAI_{+r}$ | $CAI_{+1}$ | | $CAI_{+r}$ | | $CAI_{+n}$ | CAIA |

Step 1.1 The attribute $A_1$ is age.

Step 1.2 and 1.3 Get the minimum value $d_1 = 1$ and the maximum value $d_n = 35$.

Step 1.4 Sort the age value in ascending order as shown in Fig. 2.

Step 1.5 Merge all adjacent pairs in the set that belongs to the same class, the result of merge example from Fig. 2 of interval number 1 is [1-3], interval number 2 is [4-7], and so on.

Step 1.6 to 1.7 of all n intervals, calculate the midpoint of each adjacent interval using (1).

$$B_{+r} = (Bmax_r + Bmin_{r+1})/2 \quad (1)$$

Let $B_{+r}$ be the value of interval r when $Bmax_r$ is the upper bound of interval r, and $Bmin_{r+1}$ is the lower bound of interval r+1. The example from Fig. 2, at the first interval where r = 1, $Bmax_1 = 3$, $Bmin_2 = 4$, and the midpoint $B_{+1} = (3+4)/2 = 3.5$.

Step 1.8 shows the result from Step 1. Table 2 displays the 2D-quanta matrix and discretization scheme D for attribute $A_p$ where s is the number of classes
n is the number of intervals
M is the number instances
$q_{ir}$ is the number of instances at class i and interval $(d_{r-1},d_r]$, i = 1 to s and r = 1 to n.
$M_{i+}$ is the total number of instances of class i for all intervals n.

$$M_{i+} = \sum_{r=1}^{n} q_{ir} \quad (2)$$

$M_{+r}$ is the total number of instances of individual interval $(d_{r-1},d_r]$ for all classes s.

$$M_{+r} = \sum_{i=1}^{s} q_{ir} \quad (3)$$

Table 3 shows the result of Fig. 2 with example of attribute age after Step 1 of creating 2D-quanta matrix of attribute age with 9 intervals and 3 classes.

The second step is the process to merge the intervals using class attribute interval average algorithm.

Step 2.1 is to assign the number of intervals.

Step 2.2 and 2.3 are to calculate $CAI_{+r}$ of each interval of all classes s using (4).

$$CAI_{+r} = \sum_{i=1}^{s} (q_{ir}^2/M_{i+}) * (M_{+r}) \quad (4)$$

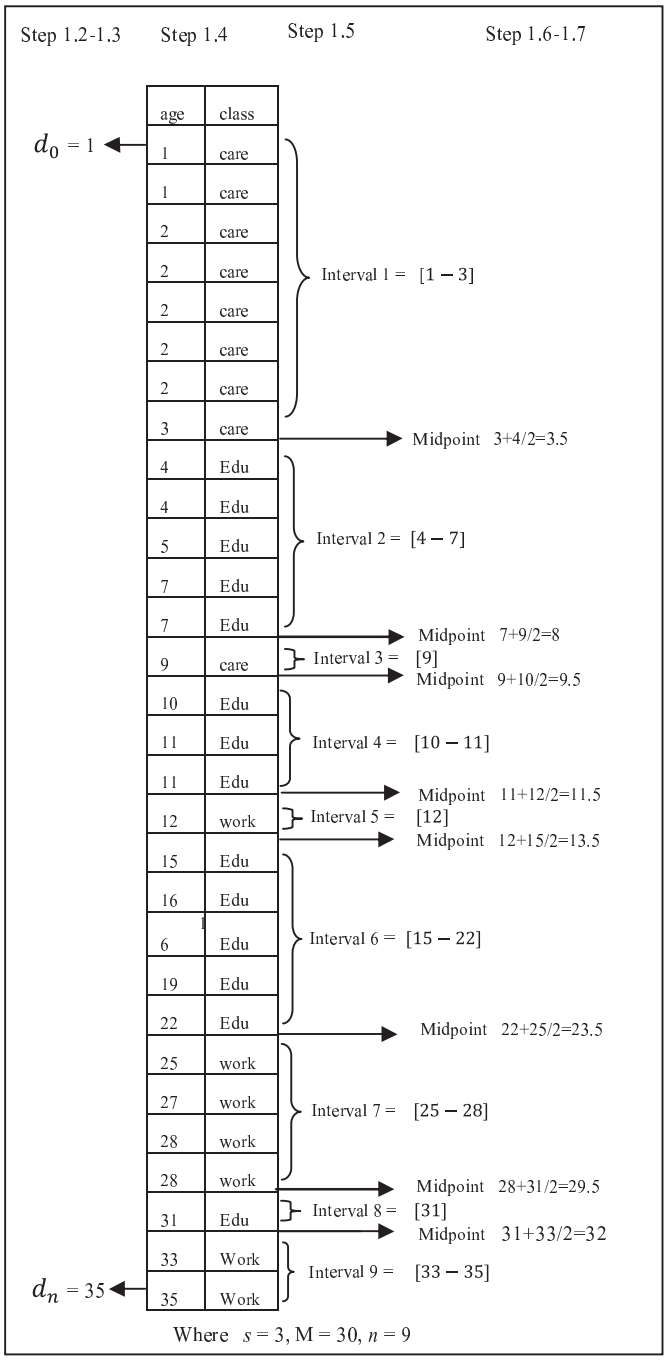



Fig. 2. Example of attribute age with 30 instances and 3 classes.

$CAI_{+r}$ value is the data distribution representative of interval r. The $CAI_{+r}$ is $q_{ir}^2$ divided by $M_{i+}$ multiple by $M_{+r}$. If the value of data distribution of interval r is low value (value of $CAI_{+r}$ is low) which means that the interval is independent with the class. In this case, the merge of this interval will have less influence to the overall output intervals. But if the value of data distribution of interval r is high (value of $CAI_{+r}$ is high), this means that interval is dependent with the class.

TABLE 3. EXAMPLE OF 2D-QUANTA MATRIX FOR AGE ATTRIBUTE.

| Class | Interval | | | | | | | | | Sum of class |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 [1 - 3.5] | 2 (3.5 - 8] | 3 (8 - 9.5] | 4 (9.5 - 11.5] | 5 (11.5 - 13.5] | 6 (13.5 - 23.5] | 7 (23.5 - 29.5] | 8 (29.5 - 32] | 9 (32 - 35] | |
| Care | 8 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| Edu | 0 | 5 | 0 | 3 | 0 | 5 | 0 | 1 | 0 | 14 |
| Work | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 2 | 7 |
| Sum of interval | 8 | 5 | 1 | 3 | 1 | 5 | 4 | 1 | 2 | 30 |
| $CAI_{+r}$ | 56.8 | 8.92 | 0.11 | 1.92 | 0.14 | 8.92 | 9.14 | **0.07** | 1.14 | **CAIA = 9.698** |

TABLE 4. THE OUTPUT DISCRETIZATION SCHEME D' OF AGE ATTRIBUTE.

| Class | Interval | | | Sum of class |
|---|---|---|---|---|
| | [1-3.5] | (3.5-23.5] | (23.5-32] | |
| Care | 8 | 1 | 0 | 9 |
| Edu | 0 | 13 | 1 | 14 |
| Work | 0 | 1 | 6 | 7 |
| **Sum of interval** | 8 | 15 | 7 | 30 |
| $CAI_{+r}$ | **56.889** | **184.881** | **36.500** | **CAIA = 92.7566138** |

Step 2.4 is the condition for stopping criteria. The algorithm will merge until the number of intervals is equal to the number of classes to ensure that the proposed algorithm will have the smallest number of intervals (equal to the number of classes).

Table 3 shows the calculation of $CAI_{+r}$ for each interval r. The $CAI_{+1}$ for interval 1 has the highest value of 56.889. The $CAI_{+8}$ of interval 8 has the lowest value of 0.071. This interval will be assigned to $CAI_{min}$ in Step 2.5.

Step 2.6 to 2.9 Candidate merge with adjacent interval on both left side and right side of interval $CAI_{min}$. The calculation for $CAIA_{left}$ and $CAIA_{right}$ are using (5).

$$CAIA = \sum_{r=1}^{n} CAI_{+r}/n \quad (5)$$

The CAIA is the value of class attribute interval average. CAIA value means the average value of data distribution for all intervals. If CAIA value is high, this means that the interval has good data distribution. Then the algorithm should select the highest value of CAIA. From Step 2.5, the $CAI_{min}$ is at interval number 8. Then the candidate merge left is interval 7 merged with interval 8. The $CAIA_{left}$ is 11.232. The candidate merge right is interval 8 merged with interval 9. The $CAIA_{right}$ is 11.00.

Step 2.10 to 2.12 is to select the new CAIA with the highest value from $CAIA_{left}$ or $CAIA_{right}$.

Step 2.13 is to reduce the number of merge intervals by one.

Step 2.14 is the output discretization scheme D' for attribute $A_p$. Table 4 shows the result of discretization scheme D' from Fig. 2. for age attribute with three intervals (the same number of classes).

## IV. EXPERIMENTAL RESULTS

The experimental design using four-benchmark UCI data sets [17]. The format of each data set (the number of instances, the number of attributes, and the number of classes) are as follows. Iris format is (150, 4, 3). Breast cancer format is (699, 10, 2). Heart disease format is (303, 13, 5). And glass format is (214, 9, 6). The proposed CAIA algorithm is compared with other six discretization algorithms that are Equal-Width (EW) [6], Equal Frequency (EF) [6], ChiMerge [7], Information Entropy Maximization (IEM) [12], Class-Attribute Interdependence Maximization (CAIM) [4], and Class-Attribute Contingency Coefficient (CACC) [2]. Four-data mining algorithms of 1) Decision Tree (J48), 2) Radial Basis Function (RBF), 3) Multilayer Perceptron (MLP), and 4) Naïve Bays (NB) are used to evaluate the performance of the discretization algorithm. The experiment is run by WEKA [18] with 10 fold cross validation.

Table 5(A) - 5(D) show comparison for the proposed CAIA algorithm with the other six discretization algorithms. The CAIA received the best accuracy mean rank for all data mining of J48, RBF, MLP, and NB with all four-data sets. The accuracy mean ranks are equal to 1.00 for J48 and RBF. The accuracy mean rank is equal to 1.25 for MLP. And the accuracy mean rank is equal to 2.25 for NB. Table 5(A) shows the achievement of J48 for the highest accuracies. CAIA received the highest accuracies for iris, breast cancer, heart disease, and glass at 98.67%, 96.57%, 90.09%, and 91.55%, respectively. Table 5(B) shows the comparison of accuracies achieved by RBF. CAIA received the highest accuracies at 98.67%, 95.99%, 82.50%, and 69.63%, respectively. Table 5(C) shows the comparison of accuracies achieved by MLP. CAIA received the highest accuracies at 98.00% for iris, 95.42% for breast cancer, and 87.45% for glass. Table 5(D) shows the comparison of accuracies achieved by NB. CAIA received the highest accuracy at 98.00% for breast cancer, and 80.85% for heart disease.

Table 6 shows the purposed CAIA and CAIM [4] with the best number of interval mean ranks at 1.75. This means that the number of intervals is small which will be good for the discretization algorithm. Because it will help the user to easily understand the data.

TABLE 5(A). THE COMPARISON OF J48 ACCURACY MEAN RANK.

| Methods | Accuracy of each data set | | | | Accuracy Mean Ranks |
|---|---|---|---|---|---|
| | Iris | Breast cancer | Heart disease | Glass | |
| EW [6] | 94.70 | 91.30 | 70.30 | 86.10 | 5.50 |
| EF [6] | 94.00 | 90.80 | 72.60 | 87.00 | 5.50 |
| ChiMerge [7] | 90.70 | 93.00 | 76.50 | 88.50 | 5.25 |
| IEM [13] | 94.00 | 93.60 | 75.20 | 89.60 | 4.00 |
| CAIM [4] | 94.00 | 93.80 | 77.10 | 90.60 | 3.00 |
| CACC [2] | 93.50 | 94.10 | 78.60 | 90.90 | 3.00 |
| **CAIA (proposed)** | 98.67 | 96.57 | 90.09 | 91.55 | 1.00 |

TABLE 5(B). THE COMPARISON OF RBF ACCURACY MEAN RANK.

| Methods | Accuracy of each data set | | | | Accuracy Mean Ranks |
|---|---|---|---|---|---|
| | Iris | Breast cancer | Heart disease | Glass | |
| EW [6] | 93.33 | 95.00 | 57.76 | 65.42 | 2.75 |
| EF [6] | 92.00 | 95.00 | 54.13 | 63.08 | 4.25 |
| ChiMerge [7] | 89.76 | 92.00 | 54.87 | 62 | 5.50 |
| IEM [13] | 94.00 | 80.98 | 58.09 | 65.7 | 3.25 |
| CAIM [4] | 92.67 | 93.42 | 50.49 | 59.81 | 5.50 |
| CACC [2] | 92.27 | 94.85 | 51.49 | 56.54 | 5.50 |
| **CAIA (proposed)** | 98.67 | 95.99 | 82.50 | 69.63 | 1.00 |

TABEL 5(C). THE COMPARISON OF MLP ACCURACY MEAN RANK.

| Methods | Accuracy of each data set | | | | Accuracy Mean Ranks |
|---|---|---|---|---|---|
| | Iris | Breast cancer | Heart disease | Glass | |
| EW [6] | 92.67 | 94.57 | 70.59 | 64.42 | 4.00 |
| EF [6] | 91.33 | 94.71 | 52.15 | 63.55 | 5.25 |
| ChiMerge [7] | 93.37 | 92.89 | 54.43 | 63.79 | 4.75 |
| IEM [13] | 93.33 | 74.69 | 55.45 | 66.17 | 4.50 |
| CAIM [4] | 94.67 | 93.56 | 46.85 | 69.16 | 3.75 |
| CACC [2] | 93.97 | 95.14 | 47.85 | 55.61 | 4.50 |
| **CAIA (proposed)** | 98.00 | 95.42 | 87.45 | 68.95 | 1.25 |

TABLE 5(D). THE COMPARISON OF NB ACCURACY MEAN RANK.

| Methods | Accuracy of each data set | | | | Accuracy Mean Ranks |
|---|---|---|---|---|---|
| | Iris | Breast cancer | Heart disease | Glass | |
| EW [6] | 93.33 | 97.28 | 57.76 | 63.55 | 3.75 |
| EF [6] | 92.67 | 96.28 | 58.75 | 69.16 | 3.75 |
| ChiMerge [7] | 93.78 | 91.88 | 59.86 | 54.23 | 4.75 |
| IEM [13] | 94.00 | 81.68 | 59.74 | 64.30 | 3.50 |
| CAIM [4] | 94.00 | 93.99 | 52.46 | 62.15 | 4.50 |
| CACC [2] | 93.34 | 95.28 | 54.46 | 59.35 | 5.25 |
| **CAIA (proposed)** | 98.00 | 96.57 | 80.85 | 61.22 | 2.25 |

TABLE 6. THE COMPARISON OF THE NUMBER OF INTERVAL MEAN RANK.

| Methods | Interval of each data set | | | | Number of Interval Mean Ranks |
|---|---|---|---|---|---|
| | Iris | Breast cancer | Heart disease | Glass | |
| EW [6] | 4.00 | 14.00 | 10.00 | 8.00 | 5.75 |
| EF [6] | 4.00 | 14.00 | 10.00 | 8.00 | 5.75 |
| ChiMerge [7] | 3.50 | 4.60 | 7.80 | 5.30 | 4.25 |
| IEM [13] | 3.00 | 2.40 | 4.00 | 2.70 | 2.25 |
| CAIM [4] | 3.00 | 2.00 | 2.00 | 6.00 | 1.75 |
| CACC [2] | 3.00 | 2.00 | 6.40 | 14.60 | 3.50 |
| **CAIA (proposed)** | 3.00 | 1.80 | 3.69 | 6.00 | 1.75 |

## V. CONCLUTION

There are two criteria to evaluate the performance of the good discretization algorithm. The first criterion is the high accuracy value. The second criterion is the small number of intervals. This paper proposed a new discretization algorithm by using supervised learning, univariate, global, static, and merge method. The Class Attribute Interval Average (CAIA) algorithm will have the smallest number of intervals. Since the stopping criteria for the algorithm is that the number of intervals always equal to the number of classes for all attributes.

The experimental results from four-data mining supervised learning algorithm, J48 decision tree, RBF and MLP neural networks, and NB, show that the proposed CAIA is out performed the six previous discretization algorithms with the best accuracy mean rank and the best number of interval mean rank. For the highest accuracy, the user will receive the best data mining model. And for the smallest number of intervals, the user will easily understand data.

## REFERENCES


[1] X. Ming and X. Xinping, "A Comparative Analysis of Discretization Algorithms for Data Mining," in Proceeding of IEEE International Conference on Grey Systems and Intelligent Services, pp. 1434-1438. November 10-12, 2009.

[2] C. J. Tsai, C. I. Lee and W. P. Yang, "A Discretization Algorithm Based on Class-Attribute Conticngency Coefficient," Information Sciences 178, pp. 174-731, 2008.

[3] L. Peng, W. Qing and G. Yujia, "Study on Comparison of Discretization Methods," in Proceeding of International Conference on Artificial Intelligence and Computational Intelligence, pp. 308-384, 2009.

[4] L. A. Kurgan and K. J. Cios, "CAIM Discretization Algorithm," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 2, pp. 145-152, February 2004.

[5] K. Shehzad, "EDISC: A Class-Tailored Discretization Technique for Rule-Based Classification," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 8, pp. 1435-1447, August 2012.

[6] A. K. C. Wong and D. K. Y. Chiu, "Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 9, pp. 796-805, 1987.

[7] R. Kerber, "ChiMerge: Discretization of Numeric Attributes," in Proceeding of the 9th National Conference on Artificial Intelligence, pp. 123-128, 1992.

[8] H. Liu and R. Setiono, "Chi2: Feature Selection and Discretization of Numeric Attribute," in Proceeding of the 7th IEEE International Conference on Tools with Artificial Intelligence, pp. 388-391, 1995.

[9] Y. Sang, K. Li and Y. Shen, "EBDA: An Effective Bottom-up Discretization Algorithm for Continuous Attributes," in Proceeding of 10th IEEE International Conference on Computer and Information Technology (CIT 2010), pp. 2455-2462, 2010.

[10] S. Garcia, J. Luengo, J. A. Saez, V. Lopez and F. Herrera, "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 4, pp. 734-750, October 2011.

[11] Y. Chaoqun, L. Jianping and D. Enming, "A Discretization Algorithm Based on Clustering and CAIR Criterion," in Proceeding of 7th IEEE International Conference on Natural Computation, pp. 1424-1429, 2011

[12] U. M. Fayyad and K. B. Irani, "Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning," in Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI), pp. 1022-1029, 1993.

[13] L. Kurgan and K. J. Cios, "Fast Class-Attribute Interdependence Maximization (CAIM) Discretization Algorithm," in Proceeding of International Conference on Machine Learning and Applications, pp. 30-36, 2003.

[14] H. Hua and H. Zhao, "A Discretization Algorithm of Continuous Attributes Based on Supervised Clustering," in Proceeding of IEEE/CCPR Conference on Pattern Recognition, pp. 1-5, 2009.

[15] S. Monti and G. F. Cooper, "A Multivariable Discretization Method for Learning Bayesian Networks from Mixed Data," in Proceeding on Uncertainly in Artificial Intelligence (UAI), pp. 404-413, 1998.

[16] Y. Kang, S. Wang, X. Liu, H. Wang and B. Miao, "An ICA-based Multivariate Discretization," in Proceeding of the First International Conference on Knowledge Science, Engineering and Management (KSEM), pp. 556-562, 2006.

[17] C. L. Blake and C. J. Merz. (1998). UCI Repository of Machine Learning Databases, University of California, Department of Information and Computer Science. [Online]. Available:www.ics.uci.edu/~mlearn/ML Repository.html (accessed 11/02/12).

[18] W. H. and E. Frank. (2005). WEKA Waikato environment for knowledge analysis, University of Waikato. [Online]. Available: www.cs.waikato.ac.nz/ml /weka/download. html (accessed 17/03/13).

# ประวัติผู้เขียน

**ชื่อ สกุล** นายอับดุลเลาะ บากา

**รหัสประจำตัวนักศึกษา** 5310220121

**วุฒิการศึกษา**

| วุฒิ | ชื่อสถาบัน | ปีที่สำเร็จการศึกษา |
|---|---|---|
| วิทยาศาสตรบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสารเพื่อการจัดการ) | มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี | 2552 |

**การตีพิมพ์เผยแพร่ผลงาน**

อับดุลเลาะ บากา, วิภาดา เวทย์ประสิทธิ และ ศิริรัตน์ วณิชโยบล. 2555. การสร้างแบบจำลองพยากรณ์นาท่วมโดยใช้เทคนิคการทาเหมืองข้อมูลของ อำเภอหาดใหญ่. Ninth Internation Joint Conference on Computer Science and Software Engineering (JCSSE 2012). กรุงเทพมหานคร ประเทศไทย. หน้า 53-58.

Baka, A., Wettayaprasit, W., and Vanichayobon, S. 2014. The Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP2014). May 6-8, 2014, Bangkok, Thailand.