วิทยานิพนธ์

การคำนวณน้ำหนักของคุณลักษณะข้อมูลสำหรับตัวจำแนกไนอีฟ เบย์ โดยวิธีการจัดกลุ่มข้อมูลและต้นไม้ตัดสินใจ

CALCULATION OF ATTRIBUTE WEIGHTS FOR NAIVE BAYES CLASSIFIER BY USING DATA CLUSTERING AND DECISION TREE

นางสาวหทัยชนก กรชี

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2551

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

ปริญญา

วิทยาการคอมพิวเตอร์	วิทยาการคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การคำนวณน้ำหนักของคุณลักษณะข้อมูลสำหรับตัวจำแนกไนอีฟ เบย์ โดยวิธีการจัดกลุ่มข้อมูลและต้นไม้ตัดสินใจ

Calculation of Attribute Weights for Naive Bayes Classifier by Using Data Clustering and Decision Tree

นามผู้วิจัย นางสาวหทัยชนก กรชี

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(รองศาสตราจารย์อนงค์นาฏ ศรีวิหค, Ph.D.)

กรรมการ .....

(ผู้ช่วยศาสตราจารย์นวลวรรณ สุนทรภิษัช, วศ.ด.)

กรรมการ .....

(รองศาสตราจารย์กฤษณะ ไวยมัย, D.U.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์ศิริกร จันทร์นวล, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่..... **เดือน**..... **พ.ศ.**.....

วิทยานิพนธ์

เรื่อง

การคำนวณน้ำหนักของคุณลักษณะข้อมูลสำหรับตัวจำแนกไนอีฟ เบย์ โดยวิธีการจัดกลุ่มข้อมูลและต้นไม้ตัดสินใจ

Calculation of Attribute Weights for Naive Bayes Classifier by Using Data Clustering and Decision Tree

โดย

นางสาวหทัยชนก กรชี

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
พ.ศ. 2551

หทัยชนก กรชี 2551: การคำนวณน้ำหนักของคุณลักษณะข้อมูลสำหรับตัวจำแนกไนอีฟเบย์ โดยวิธีการจัดกลุ่มข้อมูลและต้นไม้ตัดสินใจ ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์) สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: รองศาสตราจารย์อนงค์นาฏ ศรีวิหค, Ph.D. 92 หน้า

งานวิจัยฉบับนี้นำเสนอวิธีการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูล คือ วิธีการแบ่งกลุ่มข้อมูล (Clustering) และต้นไม้ตัดสินใจ (Decision Tree) ซึ่งปรับปรุงจากงานวิจัยของ Hall เนื่องจากค่าน้ำหนักที่ใช้ในวิธีของ Hall เป็นค่าน้ำหนักแบบสากล (Global) จึงทำให้บางกรณีมีประสิทธิภาพการทำนายไม่ดีพอ จึงมีแนวคิดในการใช้ค่าน้ำหนักแบบท้องถิ่น (Local) ซึ่งพิจารณาจากค่าน้ำหนักคุณลักษณะที่เหมาะสมกับข้อมูลทดสอบ โดยการแบ่งกลุ่มข้อมูลตามคุณลักษณะข้อมูล สำหรับวิธีการแบ่งกลุ่มที่นำเสนอในการศึกษาครั้งนี้จะใช้ 2 วิธี คือ (1) อัลกอริทึมเค-มีน (K-Means) และ (2) อัลกอริทึมสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (Cluster Feature Tree) และวิธีการลำดับชั้นแบบ Agglomerative พร้อมทั้งเปรียบเทียบประสิทธิภาพและความแม่นยำของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการแบ่งกลุ่มข้อมูลและต้นไม้ตัดสินใจกับวิธีการต้นไม้ตัดสินใจของ Hall โดยวัดค่าความถูกต้องของการทำนายและค่า Root Relative Square Error (RRSE) งานวิจัยนี้ใช้ข้อมูลจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) จำนวน 4 ชุดข้อมูล ได้แก่ (1) Chess End-Game (KR) (2) Balance Scale Weight และ Distance Database (Bal_sc) (3) Wave Form และ (4) German Credit dataset (Crd)

จากผลการวิจัยพบว่า ตัวแปรที่มีผลต่อการเพิ่มประสิทธิภาพและความแม่นยำให้แก่วิธีการคำนวณน้ำหนักคุณลักษณะจากการจัดกลุ่มและต้นไม้ตัดสินใจคือ จำนวนข้อมูลที่มีขนาดใหญ่ จำนวนคุณลักษณะของข้อมูลที่มีจำนวนมากและ จำนวนคลาสของข้อมูลที่มีจำนวนน้อย สำหรับเวลาที่ใช้ในการสร้างโมเดลของงานวิจัยนี้มีค่ามากเมื่อเปรียบเทียบกับงานวิจัยของ Hall แต่มีประสิทธิภาพการทำนายที่เพิ่มขึ้น

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่อประธานกรรมการ

Hathaichanok Kornchee 2008: Calculation of Attribute Weights for Naive Bayes Classifier by Using Data Clustering and Decision Tree. Master of Science (Computer Science), Major Field: Computer Science, Department of Computer Science. Thesis Advisor: Associate Professor Anongnart Srivihok, Ph.D. 92 pages.

The objective of this study is to improve Hall's algorithm by applying a local weighting scheme for optimizing each test data and solving problems of attribute weights which equal to zero. To improve the prediction performance, we purpose a data clustering and decision tree algorithms to calculate weights for each attribute and applied to each node of Naïve Bayes classifier. In this study two clustering methods and decision tree were used to calculate weights of Naïve Bayes. The clustering methods include (1) K-Means algorithm and (2) two step clustering by Cluster Feature Tree and Agglomerative Clustering. Then the performance of the two proposed methods were compared with Hall's method. The measures included percent of correct data prediction and Root Relative Square Error. Four Data sets used in the experiment were obtained from the University of California, Irvin (UCI) machine learning repository. They included (1) Chess End-Game (KR) (2) Balance Scale Weight and Distance Database (Bal_sc) (3) Wave Form and (4) German Credit dataset (Crd).

Result showed that data clustering and decision tree algorithms outperformed Hall's algorithm upon 3 parameters which included number of data (large), number of attribute (large) and number of class (small). However the time complexity of this approach is more than Hall's because it uses more algorithms in calculating attribute weights, clustering algorithm.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. อนงค์นาฏ ศรีวิหค ประธานกรรมการที่ปรึกษา และผู้ช่วยศาสตราจารย์ ดร. นวลวรรณ สุนทรภิษัช กรรมการวิชาเอก สำหรับความช่วยเหลือ และคำแนะนำที่มีให้ตลอดมา ขอกราบขอบพระคุณรองศาสตราจารย์ ดร.กฤษณะ ไวยมัย กรรมการวิชารอง ที่กรุณาให้ความช่วยเหลือในการสอบวิทยานิพนธ์ให้สำเร็จลุล่วงไปด้วยดี ขอกราบขอบพระคุณคณาจารย์ทุกท่าน ที่แนะนำ สั่งสอน และให้ความรู้แก่ข้าพเจ้าตลอดระยะเวลาการศึกษา

ขอขอบพระคุณ คุณพ่อ คุณแม่ ของข้าพเจ้า ที่ให้ความรัก ความห่วงใย คอยให้กำลังใจ พร้อมทั้งให้ความช่วยเหลือและสนับสนุนในด้านค่าใช้จ่ายในการศึกษาจนสำเร็จลุล่วงได้ ขอขอบคุณพี่วงกต พจน์พงศ์สรรค์ที่ให้คำแนะนำและตอบคำถามต่างๆ มากมาย ขอขอบคุณเพื่อนๆ พี่ ๆ ภาควิชาวิทยาการคอมพิวเตอร์ ทั้งรุ่นเดียวกัน รุ่นพี่ และรุ่นน้อง ที่ช่วยเหลือ และให้คำแนะนำในการทำวิทยานิพนธ์ตลอดมา รวมทั้งช่วยผลักดันการทำวิทยานิพนธ์เล่มนี้ให้สำเร็จได้ด้วยดี

สุดท้ายขอขอบคุณภาควิชาวิทยาการคอมพิวเตอร์ที่ให้โอกาสข้าพเจ้าได้เรียนรู้จนสำเร็จการศึกษา งานวิทยานิพนธ์นี้ ข้าพเจ้าหวังเป็นอย่างยิ่งว่าจะเป็นประโยชน์ต่อผู้ศึกษา ค้นคว้าและสนใจ หากข้อผิดพลาดประการใด ข้าพเจ้าขอน้อมรับไว้เพื่อนำไปใช้ในการปรับปรุงให้วิทยานิพนธ์นี้มีความสมบูรณ์ยิ่งขึ้น สำหรับความดีที่ได้รับจากวิทยานิพนธ์นี้ ข้าพเจ้ามอบแก่ผู้มีพระคุณทุกท่าน

หทัยชนก กรชี

พฤษภาคม 2551

สารบัญ

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

สารบัญตาราง (ต่อ)

ตารางที่ **หน้า**

สารบัญตาราง (ต่อ)

สารบัญภาพ

คำอธิบายสัญลักษณ์และคำย่อ

AUC	=	Area under the ROC Curve
Bal_sc	=	ชุดข้อมูลตัวอย่าง Balance Scale Weight& Distance Database
BIC	=	Schwarz's Baysian Criterion
CDW	=	การคำนวณน้ำหนักคุณลักษณะจากการจัดกลุ่มและต้นไม้ตัดสินใจ
CF Tree	=	ต้นไม้จัดกลุ่มคุณลักษณะ
Crd	=	ชุดข้อมูลตัวอย่าง German Credit dataset
HW	=	การคำนวณน้ำหนักคุณลักษณะจากต้นไม้ตัดสินใจของ Hall
KR	=	ชุดข้อมูลตัวอย่าง Chess End-Game - King & Rook
NB	=	ตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน
NBCK	=	ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มแบบเค-มีนและต้นไม้ตัดสินใจ
NBCT	=	ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะกับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ
RRSE	=	Root Relative Square Error
SD	=	ส่วนเบี่ยงเบนมาตรฐาน
WNB	=	ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก

การคำนวณน้ำหนักของคุณลักษณะข้อมูลสำหรับตัวจำแนกไนอีฟ เบย์ โดยวิธีการจัดกลุ่มข้อมูลและต้นไม้ตัดสินใจ

Calculation of Attribute Weights for Naive Bayes Classifier by Using Data Clustering and Decision Tree

คำนำ

ตัวจำแนกไนอีฟ เบย์ (Naïve Bayes Classifier) เป็นตัวจำแนกข้อมูลซึ่งมีประสิทธิภาพในงานด้านการทำนายข้อมูลที่ดี โดยใช้หลักของค่าความน่าจะเป็นของข้อมูลบนพื้นฐานของทฤษฎีเบย์ (Bayes Theorem) และสมมติฐานที่กำหนดให้การเกิดของเหตุการณ์ต่างๆ ที่ใช้ในการจำแนกนั้นเป็นอิสระต่อกัน ตัวจำแนกไนอีฟ เบย์เป็นตัวจำแนกที่นักวิจัยส่วนใหญ่ให้ความสนใจในเรื่องการทำนาย โดยจะมีการเปรียบเทียบกับวิธีการจำแนกอื่นๆ ที่ได้รับการปรับปรุงและมีการนำตัวจำแนกไนอีฟ เบย์ไปปรับปรุงและพัฒนาเพื่อให้มีประสิทธิภาพที่ดียิ่งขึ้น เนื่องจากตัวจำแนกไนอีฟ เบย์ใช้เวลาการเรียนรู้แปรตามจำนวนข้อมูลการเรียนรู้และคุณลักษณะข้อมูล

ข้อด้อยของตัวจำแนกไนอีฟ เบย์คือสมมุติฐานของความน่าจะเป็นที่คำนวณจากข้อมูลที่มีคุณลักษณะที่มีความเกี่ยวเนื่องกัน มีผลกระทบกับการคำนวณค่าความน่าจะเป็น ในการคำนวณของตัวจำแนกไนอีฟ เบย์มีการกำหนดค่าน้ำหนักของคุณลักษณะทุกตัวมีค่าเท่ากัน ปัจจุบันนักวิจัยปรับปรุงการทำงานของตัวจำแนกไนอีฟ เบย์ให้มีประสิทธิภาพและความแม่นยำในการจำแนกข้อมูลเพิ่มขึ้น โดยงานวิจัยของ Zhang และ Sheng (2004) ได้ปรับปรุงการคำนวณค่าความน่าจะเป็นของตัวจำแนกไนอีฟ เบย์โดยการนำค่าน้ำหนักของคุณลักษณะข้อมูลเข้าร่วมในการคำนวณค่าความน่าจะเป็นของข้อมูล หรือเรียกว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (Weighted Naive Bayes) ซึ่งก็ทำให้เพิ่มประสิทธิภาพและความแม่นยำ และต่อมา Hall (2007) นำเสนอการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูลโดยต้นไม้ตัดสินใจ และนำค่าน้ำหนักที่ได้นั้นไปใช้กับ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก พบว่าประสิทธิภาพของการทำนายข้อมูลนั้นดีขึ้นเมื่อเปรียบเทียบกับอัลกอริทึมการคำนวณค่าน้ำหนักคุณลักษณะอื่น ๆ คือ อัตราส่วนเกน (Gain) อัลกอริทึมรีลีฟฟ์ (ReliefF algorithm) การเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ (correlation-based feature selection) กฎการเลือกของเบย์ (Selective Bayes) ตัวจำแนกการเลือกแบบเบย์เซียน (Selective Bayesian classifier) และ NBTree (Naïve Bayes Tree)

เมื่อพิจารณางานวิจัยของ Hall แสดงให้เห็นว่าต้นไม้ตัดสินใจของ Hall (2007) มีประสิทธิภาพที่ดีกว่าตัวจำแนกอื่นดังกล่าวมาแล้วข้างต้น แต่ค่าน้ำหนักของคุณลักษณะที่ใช้เป็นค่าน้ำหนักแบบโกลบอล (ค่าน้ำหนักที่ใช้กับทุกข้อมูล) ซึ่งได้มาจากค่าส่วนกลับของรากที่สองของระดับชั้นที่น้อยที่สุดที่คุณลักษณะหนึ่งปรากฏบนต้นไม้ตัดสินใจที่สร้างจากข้อมูลการเรียนรู้ และการคำนวณน้ำหนักนี้มีปัญหาเกี่ยวกับค่าน้ำหนักคุณลักษณะบางคุณลักษณะที่ให้ค่าเป็นศูนย์ กรณีที่ปรับค่าน้ำหนักคุณลักษณะเท่ากับ 1 แทนค่าศูนย์พบว่าสามารถเพิ่มประสิทธิภาพการทำนายที่ดีขึ้น และจากงานวิจัยของ Cardie และ Howe (1997) พบว่า การใช้ค่าน้ำหนักแบบโลคอล ซึ่งเป็นค่าน้ำหนักที่พิจารณาเฉพาะแต่ละข้อมูล ทำให้ได้ประสิทธิภาพที่ดีเช่นกัน เมื่อพิจารณาค่าน้ำหนักคุณลักษณะของ Hall และ ค่าน้ำหนักคุณลักษณะของ Cardie และ Howe พบว่าได้มาจากวิธีการเรียนรู้เช่นกัน แต่กรณีของ Cardie และ Howe ต่างกับ Hall คือการเลือกค่าน้ำหนักของคุณลักษณะที่มีความเหมาะสมกับข้อมูลทดสอบจากการพิจารณาคุณลักษณะบนต้นไม้ตัดสินใจ C4.5 ด้วยเหตุนี้ในงานวิจัยนี้จึงเลือกใช้ค่าน้ำหนักแบบท้องถิ่น โดยพิจารณาจากค่าน้ำหนักคุณลักษณะที่เหมาะสมกับข้อมูลทดสอบจากวิธีการแบ่งกลุ่มข้อมูล เพื่อให้ได้ค่าน้ำหนักที่ส่งผลให้การทำนายมีประสิทธิภาพเพิ่มขึ้น โดยงานวิจัยฉบับนี้นำเสนอวิธีการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูลซึ่งปรับปรุงจากงานวิจัยของ Hall คือ วิธีการแบ่งกลุ่มข้อมูล (Clustering) และต้นไม้ตัดสินใจ (Decision Tree) สำหรับวิธีการแบ่งกลุ่มที่นำเสนอในงานวิธีนี้จะใช้ 2 วิธี คือ (1) แบบเค-มีน และ (2) แบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) และวิธีการลำดับชั้นแบบ Agglomerative พร้อมทั้งเปรียบเทียบประสิทธิภาพและความแม่นยำของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการแบ่งกลุ่มข้อมูลและต้นไม้ตัดสินใจกับวิธีการต้นไม้ตัดสินใจของ Hall โดยวัดค่าความถูกต้องของการทำนายและค่า Root Relative Square Error (RRSE) งานวิจัยนี้ใช้ข้อมูลจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) จำนวน 4 ชุดข้อมูล ได้แก่ (1) Chess End-Game (KR) (2) Balance Scale Weight และ Distance Database (Bal_sc) (3) Wave Form และ (4) German Credit dataset (Crd)

วัตถุประสงค์

1. เพื่อศึกษาทฤษฎีการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูล

2. เพื่อนำเสนอวิธีการคำนวณค่าน้ำหนักของคุณลักษณะสำหรับตัวจำแนกไนอีฟ เบย์ด้วยวิธีการจัดกลุ่ม (Clustering) และต้นไม้ตัดสินใจ (Decision Tree)

3. เพื่อเปรียบเทียบประสิทธิภาพของตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ซึ่งได้ค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) กับวิธีคำนวณน้ำหนักคุณลักษณะจากต้นไม้ตัดสินใจของ Hall

ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้สามารถเพิ่มประสิทธิภาพให้กับวิธีคำนวณน้ำหนักคุณลักษณะจากต้นไม้ตัดสินใจของ Hall (HW) โดยใช้วิธีการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) เพื่อให้ได้น้ำหนักคุณลักษณะที่เหมาะสมกับการทำนายแบบตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก

2. นำแนวคิดในการคำนวณหาน้ำหนักของคุณลักษณะโดยวิธี CDW นี้ไปใช้หาน้ำหนักของคุณลักษณะสำหรับอัลกอริทึมอื่น เช่น นิวรอลเน็ตเวิร์ก

ขอบเขตของงานวิจัย

1. ข้อมูลที่ใช้ทดสอบเป็นของมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI) machine learning repository) จำนวน 4 ชุดข้อมูล คือ Wave Form, Chess End-Game : King-Rook, German credit dataset, Balance Scale Weight& Distance Database

2. ในการทดลองจะใช้การจัดกลุ่มข้อมูล 2 วิธี คือ (1) การจัดกลุ่มแบบเค-มีน และ (2) การจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะและวิธีการลำดับชั้นแบบ Agglomerative โดยจำนวนกลุ่มที่ใช้ในการทดลอง มีขนาด 2 กลุ่มถึง 6 กลุ่ม

การตรวจเอกสาร

1. การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลคือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่เกี่ยวข้องกับชุดข้อมูลหนึ่งๆ (บุญเสริม, 2002) ในปัจจุบันมีการนำเทคนิคของการทำเหมืองข้อมูลไปประยุกต์ใช้ในงานหลายประเภท เช่น ในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร และการจัดการกับความสัมพันธ์ของลูกค้า ในด้านวิทยาศาสตร์และการแพทย์ ใช้ในการตรวจและวิจัยปัญหาบางอย่าง ด้านเศรษฐกิจและสังคม และด้านอุตสาหกรรม เพื่อลดค่าใช้จ่ายและเวลาในการจัดการกับข้อมูล

อัลกอริทึมในการทำเหมืองข้อมูล สามารถแบ่งออกเป็นสองประเภท คือ

1) การสร้างแบบจำลองในการทำนาย (Predictive modeling) ใช้ในการทำนายผลข้อมูล โดยจะมุ่งเน้นเรื่องการแบ่งแยกข้อมูลออกเป็นกลุ่มตามคุณสมบัติของคลาส (ประเภทของข้อมูล) กรณีถ้าคลาสเป็นค่าไม่ต่อเนื่อง จะเรียกกระบวนการแบ่งแยกนี้ว่า การจำแนก แต่ถ้าคลาสเป็นค่าต่อเนื่อง จะเรียกกระบวนการแบ่งแยกนี้ว่า การถดถอย (Regression)

2) การสร้างแบบจำลองในการบรรยาย เป็นการหาค่าความสัมพันธ์ต่าง ๆ (Association) หรือจัดกลุ่มข้อมูล (Clustering) โดยไม่มีจุดมุ่งหมายเพื่อการทำนาย

2. การจำแนกข้อมูล (Classification)

การจำแนกข้อมูลเป็นกระบวนการสร้างตัวจัดการข้อมูล เพื่อแสดงให้เห็นความแตกต่างระหว่างคลาส หรือกลุ่มของข้อมูล และเพื่อทำนายว่าข้อมูลนี้ควรจัดอยู่ในคลาสใด ซึ่งโมเดลที่ใช้จำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้จะขึ้นอยู่กับการวิเคราะห์ข้อมูลเรียนรู้ (Training data) (บุญเสริม, 2002) อัลกอริทึมที่ใช้ในการจำแนกข้อมูล เช่น ต้นไม้ตัดสินใจ ตัวจำแนกไนอีฟ เบย์ และนิวรอล เน็ตเวิร์ก เป็นต้น

ผลลัพธ์ที่ได้จากการเรียนรู้ คือ ตัวจำแนกข้อมูล (classifier) ซึ่งสามารถแทนได้ในหลายรูปแบบ เช่น กฎการจำแนก (If…then) ต้นไม้ตัดสินใจ ต่อมาจึงนำข้อมูลส่วนที่เหลือจากข้อมูลเรียนรู้เป็นข้อมูลทดสอบ (test data) ซึ่งเป็นกลุ่มข้อมูลที่แท้จริง ข้อมูลทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่เรียนรู้จากตัวจำแนกเพื่อทดสอบความถูกต้อง โดยเราจะปรับปรุงตัวจำแนกจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา เราจะนำข้อมูลผ่านตัวจำแนกซึ่งสามารถทำนายกลุ่มของข้อมูลนี้ได้

2.1 ต้นไม้ตัดสินใจ (Decision Tree)

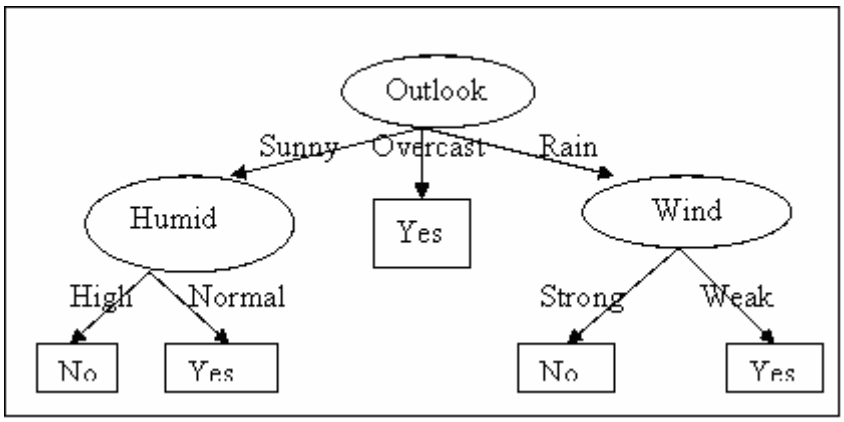
ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้ที่ใช้มากที่สุดแบบหนึ่งและเป็นการจำแนกข้อมูลให้เป็นประเภทของข้อมูล (Class) โดยพิจารณาจากค่าคุณลักษณะข้อมูล (บุญเสริม, 2002)

องค์ประกอบของต้นไม้ตัดสินใจ

1. โหนดภายใน (Internal node) คือคุณลักษณะของข้อมูล โหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้เรียกว่าโหนดราก จากภาพที่ 1 แทนด้วยวงรี เช่น คุณลักษณะ Outlook Humid และ Wind ซึ่งคุณลักษณะ Outlook เป็นโหนดเริ่มต้นจึงเป็นโหนดราก

2. กิ่ง (Branch) คือค่าที่เป็นไปได้ของแต่ละคุณลักษณะของข้อมูล ซึ่งขึ้นอยู่กับว่าเป็นคุณลักษณะใดหรือโหนดภายในใด จากภาพที่ 1 แทนด้วยลูกศร เช่น คุณลักษณะ Outlook มีค่าที่เป็นไปได้ คือ Sunny Overcast และ Rain

3. โหนดใบ (Leaf node) คือคำตอบของข้อมูลหรือผลของการจำแนกข้อมูล ส่วนมากมักเรียกว่า คลาส จากภาพที่ 1 คือ กล่องสี่เหลี่ยม Yes และ No ซึ่งเป็นคำตอบที่ต้องการของแต่ละข้อมูล



ภาพที่ 1 แสดงตัวอย่างต้นไม้ตัดสินใจ
ที่มา : Mitchell (1997)

พารามิเตอร์ที่ใช้พิจารณาเพื่อสร้างต้นไม้ตัดสินใจมีดังนี้ คือ

1. เอนโทรพี (Entropy)

ค่าเอนโทรพี คือ ค่าสารสนเทศของข้อมูล ขึ้นอยู่กับความน่าจะเป็นของข้อมูล สามารถคำนวณได้จากสมการที่ 1 ซึ่งเป็นลักษณะการวัดความไม่บริสุทธิ์หรือการกระจายของข้อมูลการเรียนรู้ ซึ่งสามารถบอกประสิทธิภาพของแต่ละคุณลักษณะข้อมูลในการจำแนกข้อมูลได้ (Mitchell, 1997) ตัวอย่างเช่น ถ้าชุดข้อมูล S ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1,m_2,...,m_n\}$ ในกรณีที่ค่าเอนโทรพีของ m_1 มีค่าเป็นศูนย์ นั่นคือความน่าจะเป็นที่จะเกิด m_1 มีค่าเป็น 1 หรือผลที่ได้มีแต่ค่า m_1 และค่าเอนโทรพีจะค่อยๆ เพิ่มขึ้นจนสูงที่สุดเมื่อความน่าจะเป็นของการเกิด m_1 เท่ากับความน่าจะเป็นของการเกิดคุณลักษณะอื่น แสดงให้เห็นว่าค่าเอนโทรพีที่น้อยจะบ่งบอกว่าชุดข้อมูลนั้นมีข้อมูลที่มีคลาสแตกต่างกันน้อยหรือเกือบจะเป็นคลาสเดียวกันหมด แต่ถ้าค่าเอนโทรพีสูงจะบ่งบอกว่ากลุ่มข้อมูลนั้นมีความแตกต่างกันมาก หรือประกอบด้วยข้อมูลหลายคลาส

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

โดยที่ $E(S)$ คือค่าเอนโทรพีของชุดข้อมูล S
$P(m_i)$ คือค่าความน่าจะเป็นที่จะเกิดค่า m_i

2. ค่าเกน (Gain)

ค่าเกน คือ ค่าที่ใช้สำหรับการตัดสินใจเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนดในต้นไม้ตัดสินใจชนิด ID3 (Quinlan, 1986) โดยการคำนวณค่าเกนของคุณลักษณะแต่ละตัวเมื่อใช้คุณลักษณะนั้นแบ่งข้อมูลเพื่อเป็นโหนดถัดไปในต้นไม้ตัดสินใจ ค่าเกนนี้คำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ นั้นคือค่าเอนโทรพี โดยที่ข้อมูล คือ S และคุณลักษณะที่เป็นโหนดคือ X โดยที่ X มีค่าทั้งหมดที่เป็นไปได้ n ค่า คือ $(v_1, v_2, ..., v_n)$ ซึ่งสามารถแบ่งข้อมูลออกเป็นกิ่งได้เป็น $\{S_{v_1}, S_{v_1}, ..., S_{v_n}\}$ จึงสามารถคำนวณค่าเอนโทรพีหลังจากแบ่งตามคุณลักษณะ X ดังสมการ

$$Gain(X) = E(S) - \sum_{i=1}^{n} \frac{|S_{v_i}|}{|S|} \times E(S_{v_i}) \tag{2}$$

โดยที่ $Gain(X)$ คือค่าเกน

$E(S)$ คือค่าเอนโทรพีของชุดข้อมูล S

S_{v_i} คือจำนวนข้อมูลที่มีค่าของคุณลักษณะ X เป็น v_i

$E(S_{v_i})$ คือค่าเอนโทรพีของคุณลักษณะ X เป็น v_i

S คือจำนวนข้อมูลทั้งหมด

3. ค่าสารสนเทศของการแบ่งแยก (Split Information)

ค่าสารสนเทศของการแบ่งแยกของคุณลักษณะแสดงถึงระดับการกระจายของข้อมูล สามารถคำนวณได้จากสมการดังนี้

$$Split\_inFo(X) = \sum_{i=1}^{n} \frac{|S_{v_i}|}{|S|} log_2 \frac{|S_{v_i}|}{|S|} \tag{3}$$

โดยที่ $Split\_inFo(X)$ คือค่าสารสนเทศของการแบ่งแยกของคุณลักษณะ X

S_{v_i} คือจำนวนข้อมูลที่มีค่าของคุณลักษณะ X เป็น v_i

S คือจำนวนข้อมูลทั้งหมด

4. อัตราส่วนเกน (Gain Ratio)

อัตราส่วนเกน เป็นหลักในการเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนดในต้นไม้ตัดสินใจชนิด C4.5 (Quinlan, 1993) เนื่องจากหากใช้ค่าเกนในการคัดเลือกคุณลักษณะจะมีอคติ (bias) อย่างมากกับข้อมูลที่ประกอบด้วยคุณลักษณะที่มีค่าที่เป็นไปได้จำนวนมาก ๆ เช่น ข้อมูลเลขประจำตัว ซึ่งเมื่อคำนวณค่าเกนแล้วคุณลักษณะนี้จะถูกเลือกเนื่องจากค่าเอนโทรพีของคุณลักษณะนี้จะได้เท่ากับศูนย์ ทำให้ค่าเกนที่ได้มีค่าสูงที่สุดเสมอ การแก้ไขความอคติทำได้โดยการปรับค่าเกนให้ถูกต้องโดยใช้ค่าสารสนเทศของการแบ่งแยกไปหารค่าเกน จึงได้ค่าอัตราส่วนเกนของแต่ละคุณลักษณะ ซึ่งช่วยแก้ไขความอคติของค่าเกนได้ ค่าอัตราส่วนเกน คำนวณได้จากสมการที่ 4

$$Gain\_ratio(X) = \frac{Gain(X)}{split\_inFo(X)} \qquad (4)$$

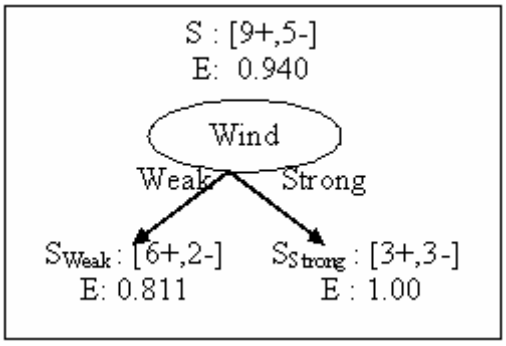
โดยที่ *Gain_ratio(X)* คือค่าอัตราส่วนเกนของคุณลักษณะ *X*

Gain(X) คือค่าเกนของคุณลักษณะ *X*

Split_inFo(X) คือค่าสารสนเทศของการแบ่งแยกของคุณลักษณะ *X*

ตัวอย่างแสดงวิธีการคำนวณค่าอัตราส่วนเกน

พิจารณาจากภาพที่ 2 คำนวณหาค่าอัตราส่วนเกนของคุณลักษณะ Wind



ภาพที่ 2 แสดงโหนดคุณลักษณะ Wind เพื่อหาค่าอัตราส่วนเกน

ที่มา : Mitchell, 1997

Gain(S,Wind) = 0.940 – ((8/14) x 0.811) – ((6/14) x 1.00) = 0.048

Split_inFo(Wind) = - (8/14) log2(8/14) - (6/14) log2(6/14) = 0.46+0.523 = 0.983

Gain_ratio(Wind) = 0.048/0.983 = 0.049

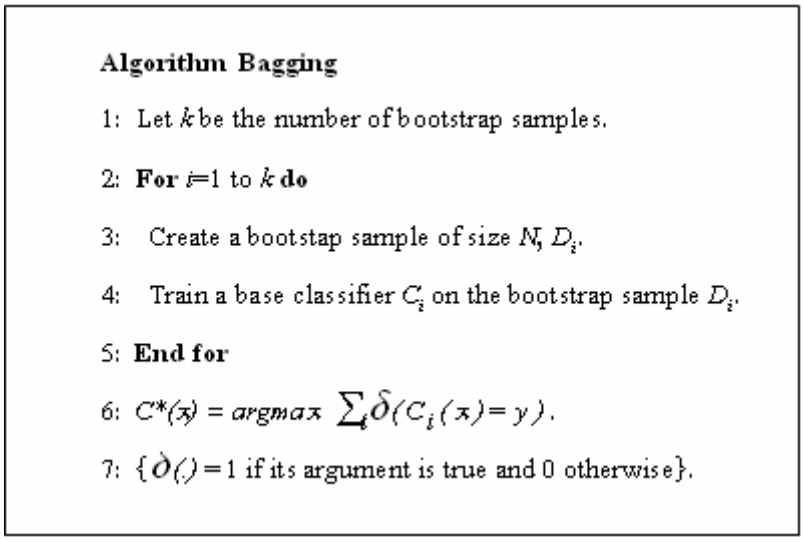
สำหรับต้นไม้ตัดสินใจมีความซับซ้อนด้านเวลาในการเรียนรู้คือ $O(n^2*N)$ (Su and Zhang, 2006) โดยที่ N คือจำนวนข้อมูลการเรียนรู้ และ n คือจำนวนคุณลักษณะของข้อมูล สำหรับกรณีการทดสอบข้อมูลจะมีความซับซ้อนด้านเวลา คือ O(N*D) โดยที่ D แทนระดับความลึกสุดท้ายของต้นไม้ตัดสินใจ (Witten, 2000)

2.2 วิธีการแบ็คกิ้ง (Bagging Method)

การรวมกระบวนการ (Ensemble) มีแนวคิดคือการสร้างตัวจำแนกหลายตัวจากข้อมูลต้นแบบเดียวกันและรวมผลการทำนายสำหรับใช้ในการจำแนกข้อมูลที่ไม่รู้จัก การรวมกระบวนการนั้นเหมาะสมกับตัวจำแนกที่ไม่เสถียร เช่น ต้นไม้ตัดสินใจ ตัวจำแนกกฎพื้นฐาน และโครงข่ายประสาทเทียม (Tan *et al*., 2006)

วิธีการแบ็คกิ้ง เป็นวิธีที่ใช้เทคนิคการรวมกระบวนการ (Ensemble) อีกวิธีหนึ่ง ซึ่งใช้วิธีการสุ่มข้อมูลแบบแทนที่ (Sampling with replacement) ในปริมาณที่เท่ากันในแต่ละรอบ ใน 1 ข้อมูลอาจจะพบได้ในข้อมูลการเรียนรู้เดียวกัน และก็อาจมีบางข้อมูลที่ไม่ปรากฏในชุดข้อมูลการเรียนรู้ชุดใดเลยก็ได้

วิธีการแบ็คกิ้ง ช่วยลดความแปรผันของตัวจำแนกโดยประสิทธิภาพของวิธีการแบ็คกิ้งขึ้นอยู่กับความเสถียรของตัวจำแนก ถ้าหากว่าตัวจำแนกไม่เสถียร วิธีการแบ็คกิ้ง จะช่วยลดความคลาดเคลื่อนที่เกี่ยวกับความไม่คงที่ในการสุ่มของชุดข้อมูลการเรียนรู้ แต่ถ้าหากว่าตัวจำแนกมีความเสถียรแล้วค่าความคลาดเคลื่อนของการรวมกระบวนการจะเป็นค่าอคติ (Bias) ของตัวจำแนก ในสถานการณ์นี้ วิธีการแบ็คกิ้ง จะไม่สามารถปรับปรุงประสิทธิภาพของตัวจำแนกได้อย่างมีนัยสำคัญ ซึ่งอาจจะส่งผลลดประสิทธิภาพตัวจำแนกด้วย วิธีการแบ็คกิ้ง มีการทำงานตามอัลกอริทึม ดังภาพที่ 3



ภาพที่ 3 แสดงขั้นตอนการทำงานของขั้นตอนวิธีการแบ็คกิ้ง
ที่มา : Tan *et al*., 2006

ขั้นตอนวิธีการแบ็คกิ้ง (Tan et al.,2006) มีดังนี้

1) กำหนดค่า k แทนค่าจำนวนรอบของการเรียนรู้ด้วยวิธีแบ็คกิ้ง
2) ในแต่ละรอบ $i = 1$ ถึง $i = k$
3) สุ่มข้อมูลแบบแทนที่จำนวน N ข้อมูล ให้เป็นเซตของ D_i
4) นำข้อมูลในข้อ 3 ไปเรียนรู้ด้วยตัวจำแนก C_i
5) คำนวณผลการทำนายของทุกรอบ โดยที่
   5.1 ถ้าข้อมูลมีค่าต่อเนื่อง ผลการทำนายคือค่าเฉลี่ยของ k รอบ
   5.2 ถ้าข้อมูลมีค่าไม่ต่อเนื่อง ผลการทำนายคือผลโหวตที่มากที่สุดใน k รอบ

2.3 ตัวจำแนกไนอีฟ เบย์ (Naive Bayes Classifier)

ตัวจำแนกไนอีฟ เบย์ เป็นการจำแนกข้อมูลที่ใช้หลักบนความน่าจะเป็นของข้อมูลบนพื้นฐานของทฤษฎีเบย์ (Bayes Theorem) และสมมติฐานที่กำหนดให้การเกิดของเหตุการณ์ต่างๆ ที่ใช้ในการจำแนกนั้นเป็นอิสระต่อกัน ตัวจำแนกไนอีฟ เบย์เป็นตัวจำแนกข้อมูลที่มีประสิทธิภาพสำหรับงานด้านการทำนายข้อมูล ตัวจำแนกไนอีฟ เบย์ (Irina, 2001) มีสมการดังนี้

$$P(C_L|a_1,a_2,...,a_n) = P(C_L)\prod_{i=1}^{n} P(a_i|C_L) \quad (5)$$

โดยที่ $P(C_L)$ คือค่าความน่าจะเป็นของข้อมูลที่ให้คลาส

$P(a_i|C_L)$ คือค่าความน่าจะเป็นของข้อมูลคุณลักษณะที่ i มีค่า a_i และให้คลาส C_L

$P(C_L|a_1,a_2,...,a_n)$ คือค่าความน่าจะเป็นของข้อมูลทดสอบ (ที่มีค่าคุณลักษณะ $(a_1,a_2,...,a_n)$ ที่จะให้คลาส C_L

ตัวจำแนกไนอีฟ เบย์ทำงานได้ดีเมื่อทดสอบกับข้อมูลจริง และเมื่อมีการกำจัดคุณลักษณะที่ขึ้นต่อกันออกไป แต่ข้อเสียของตัวจำแนกไนอีฟ เบย์ คือ กรณีมีค่าคุณลักษณะหนึ่งไม่เกิดขึ้นในข้อมูลการเรียนรู้เลย จะทำให้ผลการทำนายไม่ถูกต้อง เพราะ $P(C_L|a_1,a_2,...,a_n)$ ในสมการที่ 5 จะมีค่าเป็นศูนย์ แต่มีวิธีที่สามารถแก้ไขโดยการปรับค่าคุณสมบัตินั้นเล็กน้อย โดยใช้วิธีการประมาณค่า m ของความน่าจะเป็น (m-estimate of probability) (Mitchell, 1997) ตามสมการที่ 6 เพื่อแก้ไขผลคูณเป็นศูนย์

$$P(a_i|C_L) = \frac{N_{a_i} + (m \times p)}{N_{C_L} + m} \quad (6)$$

โดยที่ N_{C_L} คือจำนวนของข้อมูลเรียนรู้ที่ให้คลาส C_L

N_{a_i} คือจำนวนของข้อมูลเรียนรู้ที่ให้คลาส C_L และมีคุณลักษณะที่ i มีค่า a_i

m คือค่าคงที่ หรือเรียกว่า equivalent sample size โดยมีค่าสัมพันธ์กับค่า p

p คือความน่าจะเป็นของลำดับความสำคัญ (prior estimate of the probability)

ตัวจำแนกไนอีฟ เบย์มีความซับซ้อนของเวลา คือ O(n*N) โดยที่ N คือจำนวนข้อมูลการเรียนรู้และ n คือจำนวนคุณลักษณะข้อมูล (John and Langley, 1995)

2.4 ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (Weighted Naïve Bayes)

ตัวจำแนกไนอีฟ เบย์เป็นตัวจำแนกข้อมูลที่นิยมประยุกต์ใช้กับงานหลากหลาย จึงมีการพัฒนาตัวจำแนกไนอีฟ เบย์เพื่อปรับปรุงประสิทธิภาพให้เพิ่มขึ้น การเพิ่มค่าน้ำหนักเข้าไปใน

$$P(C_L|a_1,a_2,...,a_n) = P(C_L)\prod_{i=1}^{n} P(a_i|C_L)$$

$$P(C_L|a_1,a_2,...,a_n) = P(C_L)\prod_{i=1}^{n} P(a_i|C_L)^{w_i} \quad (7)$$

โดยที่ $P(C_L)$ คือค่าความน่าจะเป็นของข้อมูลที่ให้คลาส

$P(a_i|C_L)$ คือค่าความน่าจะเป็นของข้อมูลคุณลักษณะที่ I มีค่า a_i และให้คลาส C_L

$P(C_L|a_1,a_2,...,a_n)$ คือค่าความน่าจะเป็นของข้อมูลทดสอบ (ที่มีค่าคุณลักษณะ $(a_1,a_2,...,a_n)$ ที่จะให้คลาส C_L

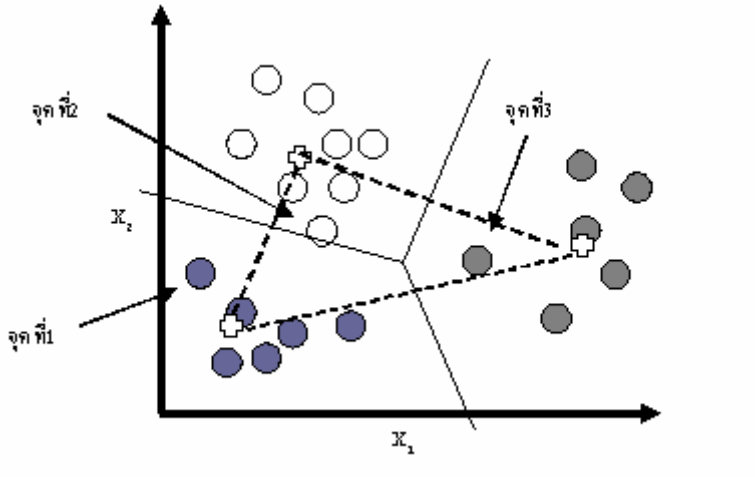
w_i คือค่าน้ำหนักของแต่ละคุณลักษณะข้อมูล

3. การจัดกลุ่ม (Clustering)

การจัดกลุ่มข้อมูล คือ การรวมกลุ่มของข้อมูลที่มีลักษณะเหมือนกัน รูปแบบและแนวโน้มที่เหมือนกัน โดยเริ่มจากการหาตัวแทนของกลุ่ม จากนั้นทำการเปรียบเทียบข้อมูลกับตัวแทนของแต่ละกลุ่ม ถ้าข้อมูลคล้ายคลึงกับตัวแทนของกลุ่มไหนก็จะถูกจัดให้อยู่กลุ่มนั้น วิธีในการแบ่งกลุ่มข้อมูลแบ่งออกเป็น 2 ประเภท (Jain *et al.*, 1999) คือ

3.1 การจัดกลุ่มแบบแบ่งส่วน (Partitional Clustering)

การจัดกลุ่มข้อมูลแบบแบ่งส่วน (Partitional Clustering) การจัดจำแนกข้อมูลออกเป็นกลุ่มย่อย ๆ ตามจำนวนกลุ่มที่กำหนด โดยไม่แสดงถึงความสัมพันธ์ระหว่างกลุ่มแต่ละกลุ่มในเชิงโครงสร้างเทคนิคการจัดกลุ่มข้อมูลแบบแบ่งส่วน ได้แก่ อัลกอริทึมเค-มีน และ Fuzzy C-Means



ภาพที่ 4 ภาพแสดงการจัดกลุ่มข้อมูลแบบแบ่งส่วน

ที่มา : Jain *et al*. (1999)

3.1.1 การจัดกลุ่มแบบ เค-มีน (K-Means Algorithm)

เทคนิคขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน (K-Means Algorithm) นั้นมีการพัฒนาและนำเสนอโดย Mac Queen ในปี 1967 ซึ่งแสดงให้เห็นถึงอัลกอริทึมในการจัดกลุ่มที่สมาชิกภายในกลุ่ม จะมีระยะใกล้จุดศูนย์กลางหรือตัวแทนของกลุ่ม (Mean) โดยขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบ เค-มีน นั้นจะประกอบด้วย การกำหนดจำนวนกลุ่มเริ่มต้น กำหนดตัวแทนกลุ่ม การจัดข้อมูลแต่ละตัวเข้ากลุ่ม และสุดท้ายคือ การปรับปรุงตัวแทนกลุ่มในแต่ละกลุ่ม ปัญหาที่พบ คือ ปัญหาการกำหนดจำนวนจุดเริ่มต้น ซึ่งส่งผลต่อประสิทธิภาพของการจัดกลุ่ม หรือการที่กลุ่มบางกลุ่มนั้นมีจำนวนสมาชิกน้อยเกินหรืออาจจะไม่มีสมาชิกในกลุ่มเลย จึงได้มีการกำหนดกฎเกณฑ์ว่ากลุ่มที่จะอยู่ในรอบถัดได้ไปนั้น จะต้องมีสมาชิกอยู่ไม่น้อยกว่าค่าคงที่ค่าหนึ่งที่กำหนดขึ้นมา

Algorithm Basic *K*-Means algorithm.

1: Select *K* points as initial centroids.

2: **repeat**

3: From *K* clusters by assigning each point to its closest centroid.

4: Recompute the centroid of each cluster.

5: **until** Centroids do not change.

ภาพที่ 5 แสดงขั้นตอนการทำงานของขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน

ที่มา : (Tan *et al*., 2006)

ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน (Tan *et al*., 2006) มีขั้นตอนดังนี้

1) กำหนดจำนวนกลุ่มที่ต้องการ โดยกำหนดให้ K เท่ากับจำนวนกลุ่มที่ต้องการแบ่งกลุ่ม และเลือกจุดศูนย์กลางของกลุ่มเริ่มต้น

2) คำนวณหาตัวแทนกลุ่ม หรือจุดศูนย์กลางของกลุ่ม (Centroid) ในแต่ละกลุ่ม จัดกลุ่มข้อมูลใหม่โดยพิจารณาจากค่าความใกล้ชิดหรือระยะห่างของข้อมูลในกลุ่มกับตัวแทนของกลุ่มต่างๆ ว่าข้อมูลนั้นสมควรที่จะอยู่กลุ่มใด

3) ทำการปรับปรุงการจัดกลุ่มโดยย้อนกลับไปทำข้อ 2 และจะหยุดเมื่อข้อมูลสมาชิกในกลุ่มแต่ละกลุ่มนั้นไม่มีการเปลี่ยนแปลงแล้ว

ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีนต้องการพื้นที่ (space) ที่ใช้ในการทำงานน้อย (modest) เนื่องจากเก็บเฉพาะข้อมูล (data points) และตัวแทนของกลุ่ม (centroids) เท่านั้น ดังนั้นจึงต้องการพื้นที่ (space complexity) เป็น O((n + K)*N) เมื่อ N คือจำนวนข้อมูลทั้งหมด และ n คือจำนวนของคุณลักษณะข้อมูล ส่วนเวลาที่ใช้ในการประมวลผลจะเป็นสมการเชิงเส้น (linear) ในรูปของจำนวนข้อมูล ดังนั้นเวลา (time complexity) ที่ต้องการใช้เป็น O(I * K * n * N) เมื่อ I เป็นจำนวนรอบที่ใช้เมื่อตัวแทนของกลุ่มมีการเปลี่ยนแปลง และ K คือ จำนวนกลุ่มที่ต้องการจัดกลุ่ม เมื่อเรากำหนดค่าตัวแปรให้คงที่ เวลาที่ใช้จะเป็น O(n*N) (Tan *et al.*, 2006)

ข้อเด่นของขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน

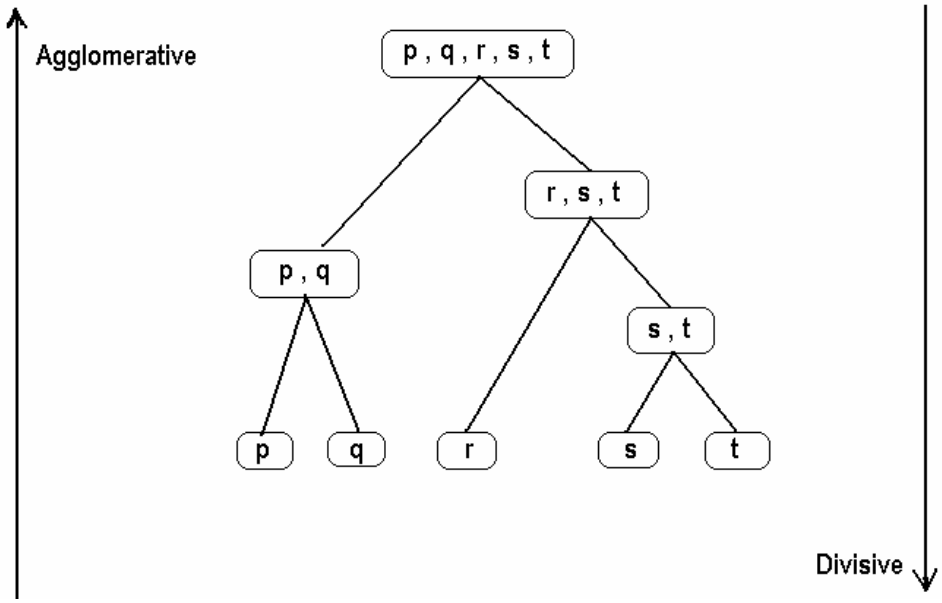
1) ง่ายและนิยมนำไปใช้งาน
2) จัดกลุ่มข้อมูลได้รวดเร็ว

ข้อด้อยของขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน

1) การกำหนดจำนวนกลุ่มและตัวแทนกลุ่มเริ่มต้น มีผลต่อประสิทธิภาพการจัดกลุ่มทั้งในเชิงเวลาและความถูกต้อง
2) ขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน จะมีปัญหากับข้อมูลที่มี outlier

3.2 การจัดกลุ่มข้อมูลแบบลำดับชั้น (Hierarchical Clustering)

ทฤษฎีการจัดกลุ่มลำดับขั้น เป็นทฤษฎีที่นิยมมากที่สุดสำหรับนำไปประยุกต์ใช้ในการจัดกลุ่มข้อมูลที่มีคุณลักษณะซ้อนและข้อมูลที่มีลักษณะค่อนข้างซับซ้อน (Tan *et al*,2006) การจัดกลุ่มแบบลำดับชั้น สามารถแบ่งออกเป็น 2 ประเภท คือ



ภาพที่ 6 แสดงการจัดกลุ่มข้อมูลแบบลำดับชั้น
ที่มา : Resampling Stat (2008)

1. Agglomerative (Bottom-Up) มีหลักการในการทำงาน คือ เริ่มจากการจัดกลุ่มของข้อมูลออกเป็นจำนวน n กลุ่มจากข้อมูลทั้งหมด n ตัว ให้เป็นบัพภายนอก (External node) แล้วทำการรวมข้อมูลที่มีค่าความเหมือนใกล้เคียงกันมากที่สุดครั้งละ 2 ตัว ทำซ้ำจนกระทั่งได้ข้อมูลสุดท้าย 1 กลุ่มซึ่งเป็นบัพราก (Root node) วิธีการคำนวณของการจัดกลุ่มแบบลำดับขั้นแบบ Agglomerative มีวิธีการ 4 ขั้นตอน ดังนี้ (Ward, 1963)

1) แบ่งข้อมูลออกเป็น n กลุ่ม (cluster) โดยในแต่ละกลุ่ม มีสมาชิก 1 ตัว

$$L = S_1, S_2, S_3, S_4, ..., S_{n-1}, S_n \tag{8}$$

โดยที่ S คือข้อมูลที่ใช้ในการจัดกลุ่มข้อมูล (Input)

n คือจำนวนข้อมูลทั้งหมด

L คือจำนวนกลุ่มของข้อมูล

2) คำนวณหาค่าความเหมือนของข้อมูลทีละคู่ (S_i, S_j) ทุกๆ ข้อมูลในเซต L โดยการคำนวณหาค่าความเหมือนสามารถคำนวณจากสมการต่อไปนี้

2.1) Single Linkage คือ คำนวณความเหมือนโดยการหาระยะทางระหว่างข้อมูล 2 กลุ่มที่ใกล้กันที่สุดซึ่งคำนวณได้จากสมการ

$$d(r,s) = \min(dist(x_{ri}, x_{si})) \tag{9}$$

โดยที่ d คือ ระยะทางคำนวณได้โดยวิธีการหาระยะทางแบบ Euclidian

r คือ กลุ่มข้อมูลที่ 1

s คือ กลุ่มข้อมูลที่ 2

$dist$ คือ ค่าของระยะทางระหว่าง 2 จุดที่หา

x_{ri} คือ ค่าของข้อมูลกลุ่มที่ 1 ตำแหน่งที่ i

x_{si} คือ ค่าของข้อมูลกลุ่มที่ 2 ตำแหน่งที่ i

2.2) Complete Linkage วิธีนี้จะตรงข้ามกับวิธี Single Linkage นั่นคือ การหาระยะทาง ระหว่างข้อมูล 2 กลุ่มที่ห่างกันที่สุดแต่วิธีนี้ไม่เป็นที่นิยมเนื่องจากมีสิ่งรบกวนมากทำให้การคำนวณใช้เวลานาน ซึ่งคำนวณได้จากสมการ

$$d(r,s) = \max(dist(x_{ri}, x_{si})) \tag{10}$$

โดยที่ d คือ ระยะทางคำนวณได้โดยวิธีการหาระยะทางแบบ Euclidian

r คือ กลุ่มข้อมูลที่ 1

s คือ กลุ่มข้อมูลที่ 2

$dist$ คือ ค่าของระยะทางระหว่าง 2 จุดที่หา

x_{ri} คือ ค่าของข้อมูลกลุ่มที่ 1 ตำแหน่งที่ i

x_{si} คือ ค่าของข้อมูลกลุ่มที่ 2 ตำแหน่งที่ i

2.3) Average Linkage เป็นวิธีการที่ผสมระหว่าง 2 วิธีแรก คือ Single Linkage กับ Complete Linkage แล้วเลือกค่าเฉลี่ยของระยะห่างระหว่าง 2 คลัสเตอร์ ซึ่งคำนวณได้จากสมการ

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \tag{11}$$

โดยที่ d คือ ระยะทางคำนวณได้โดยวิธีการหาระยะทางแบบ Euclidian

r คือ กลุ่มข้อมูลที่ 1

s คือ กลุ่มข้อมูลที่ 2

$dist$ คือ ค่าของระยะทางระหว่าง 2 จุดที่หา

x_{ri} คือ ค่าของข้อมูลกลุ่มที่ 1 ตำแหน่งที่ i

x_{sj} คือ ค่าของข้อมูลกลุ่มที่ 2 ตำแหน่งที่ j

n_r คือ จำนวนข้อมูลของกลุ่มที่ 1

n_s คือ จำนวนข้อมูลของกลุ่มที่ 2

3) ทำการรวม (Merge) คู่ของข้อมูลที่คำนวณได้จากขั้นตอนที่ 2 (S_i,S_j) และทำการสร้างต้นไม้ โดยที่โหนด S_{ij} จะเป็น Parent node ให้กับโหนด S_i และ S_j โดยใช้วิธีการรวมจากสมการใดสมการหนึ่งของสมการหาค่าความเหมือนในข้อ 2

4) กลับไปทำซ้ำขั้นตอนที่ 2 จนกว่าจะเหลือตามจำนวนคลัสเตอร์ที่กำหนดไว้

สำหรับการจัดกลุ่มลำดับชั้นแบบ Agglomerative มีความซับซ้อนด้านเวลาในการเรียนรู้คือ $O((n*N)^2)$ โดยที่ N คือจำนวนข้อมูลการเรียนรู้และ n คือ จำนวนคุณลักษณะ (Tan *et al.*, 2006)

2. Divisive (Top-Down) มีหลักการในการทำงาน เหมือนวิธีของ Agglomerative แต่ต่างกันตรงที่วิธี Divisive นี้จะทำงานจากโหนดรากไปสู่โหนดภายนอกหรือจากบนลงล่างนั่นเอง แต่วิธีนี้ไม่เป็นที่นิยมเนื่องจากใช้เวลาในการคำนวณนานกว่าแบบ Agglomerative

4. การจัดกลุ่มแบบสองขั้นตอน (Two-Step Clustering)

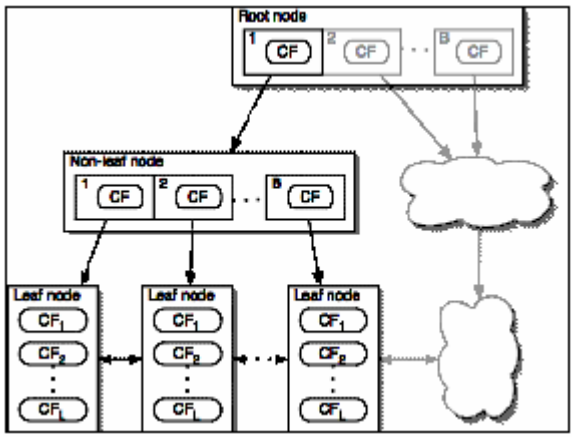
การจัดกลุ่มแบบสองขั้นตอนเป็นอัลกอริทึมที่มีการพัฒนาเพื่อเพิ่มประสิทธิภาพให้กับการจัดกลุ่ม โดยจะใช้วิธีการจัดกลุ่ม 2 แบบมาทำงานร่วมกัน ยกตัวอย่างเช่น

การจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative ซึ่งเป็นวิธีการจัดกลุ่มที่สามารถรองรับข้อมูลจำนวนมากได้อย่างมีประสิทธิภาพและรวดเร็ว โดยการทำงานนั้นแบ่งออกเป็น 2 ขั้นตอน คือ

1) Pre-cluster คือ การจัดกลุ่มข้อมูลออกเป็นกลุ่มย่อยจำนวนหลายๆ กลุ่ม โดยในขั้นตอนนี้ใช้เทคนิคต้นไม้การจัดกลุ่มตามคุณลักษณะ (Cluster feature Tree) ซึ่งเป็นเทคนิคหนึ่งของการจัดกลุ่มแบบลำดับโดยใช้หลักการพิจารณาแต่ละระเบียน (record) ของข้อมูลว่าควรที่จะรวมกับกลุ่มแรก หรือเริ่มเป็นกลุ่มใหม่โดยพิจารณาจากระยะห่างระหว่างข้อมูล โดยที่ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF tree) ประกอบไปด้วยลำดับชั้นของโหนด โดยที่โหนดใบเป็นโหนดที่ใช้แสดงกลุ่มย่อยที่ต้องการ และโหนดที่ไม่ใช่โหนดใบจะเป็นตัวแนะทางให้แก่โหนดใหม่เพื่อให้เข้าสู่โหนดใบที่ถูกต้องอย่างรวดเร็ว

เทคนิคต้นไม้การจัดกลุ่มตามคุณลักษณะ (Cluster feature Tree – CF Tree)

CF-Tree คือ ต้นไม้ที่ใช้ในการจัดเก็บข้อมูลเป็นกลุ่มย่อย (Sub-cluster) สำหรับการจัดกลุ่มข้อมูลโดยจะพิจารณาแต่ละระเบียนข้อมูล (Record) ว่าควรที่จะอยู่ในกลุ่มย่อยใดหรือจะรวมกับกลุ่มย่อยเดิม หรือเริ่มจัดเป็นกลุ่มย่อยใหม่โดยพิจารณาจากระยะห่างระหว่างข้อมูล ตัวอย่างของ CF Tree ได้แก่ อัลกอริทึม BIRCH (Zhang *et al*., 1996)



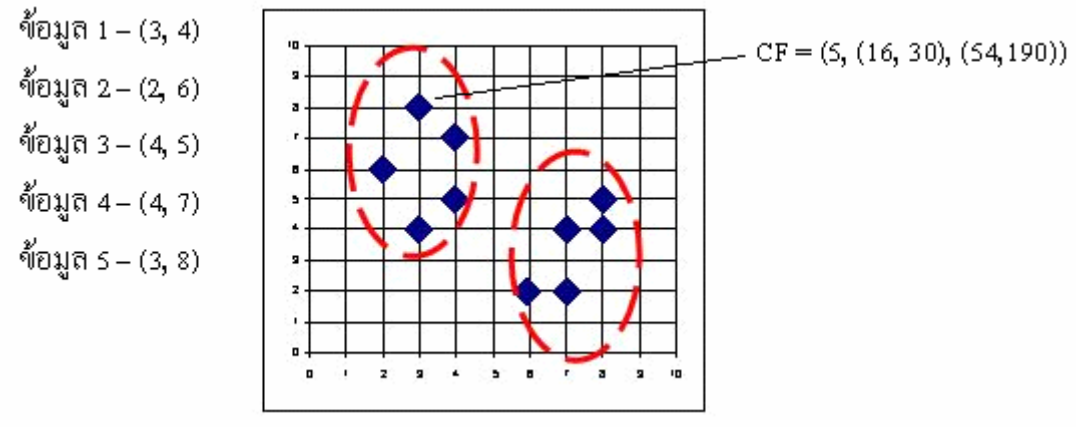
ภาพที่ 7 ลักษณะของ Clustering Feature Tree (CF-Tree)
ที่มา : Zhang *et al.* (1996)

$$CF = (N, \vec{LS}, SS) \tag{12}$$

โดยที่ N คือ จำนวนข้อมูลทั้งหมดในกลุ่มข้อมูล

$\vec{LS} = \sum_{i=1}^{N} \vec{X}_i$ คือ ผลรวมของข้อมูลในแต่ละมิติ

$SS = \sum_{i=1}^{N} \vec{X}_i^2$ คือ ผลรวมของกำลังสองของข้อมูลในแต่ละมิติ



ภาพที่ 8 ลักษณะของการจัดกลุ่มโดย Cluster Feature
ที่มา : Zhang *et al.* (1996)

สำหรับ CF-Tree มีความซับซ้อนด้านเวลาในการเรียนรู้คือ $O((n*N)^2)$ โดยที่ N คือ จำนวนข้อมูลการเรียนรู้และ n คือ จำนวนคุณลักษณะ (Zhang *et al.*,1996)

2) Cluster คือ การจัดกลุ่มใหม่อีกครั้งตามจำนวนกลุ่มที่ต้องการ โดยใช้ข้อมูลกลุ่มย่อยจากข้อ 1 ซึ่งเทคนิคที่ใช้ในขั้นตอนนี้คือ การจัดกลุ่มวิธีลำดับชั้นแบบ Agglomerative

ในปัจจุบันได้มีการกำหนดจำนวนกลุ่มที่เหมาะสม โดยพิจารณาจากค่า BIC (Schwarz's Baysian Criterion) (Fraley และ Raftery, 1998) ซึ่งสามารถคำนวณได้จากสมการดังนี้

$$BIC(J) = 2\sum_{j=1}^{J}\xi_j + m_j \log(N) \tag{13}$$

$$m_j = J \times (2 \times K^A + \sum_{k=1}^{K^B}(L_k - 1)) \tag{14}$$

$$\xi_j = -N_j \times (\sum_{k=1}^{K^A}\frac{1}{2}\log(\hat{\sigma}_k^2 + \hat{\sigma}_{jk}^2) + \sum_{k=1}^{K^B}\hat{E}_{jk}) \tag{15}$$

$$\hat{E}_{jk} = \sum_{l=1}^{L_k} \frac{N_{jkl}}{N_j} \log \frac{N_{jkl}}{N_j} \quad (16)$$

โดยที่ K^A คือค่าจำนวนตัวแปรที่เป็นค่าต่อเนื่อง

N คือจำนวนข้อมูลทั้งหมด

K^B คือค่าจำนวนตัวแปรที่เป็นค่าไม่ต่อเนื่อง

N_k คือจำนวนของข้อมูลที่อยู่ในคลัสเตอร์ k

N_{jkl} คือจำนวนของข้อมูลที่อยู่คลัสเตอร์ j และมีตัวแปรที่ต่อเนื่อง k ที่มีค่า l

L_k คือจำนวนค่าที่เป็นไปได้ของตัวแปรที่ไม่ต่อเนื่องตัวที่ k

$\hat{\sigma}_k^2$ คือค่าความแปรปรวนของตัวแปรที่ไม่ต่อเนื่องตัวที่ k สำหรับทุกข้อมูล

$\hat{\sigma}_{jk}^2$ คือค่าความแปรปรวนของตัวแปรที่ไม่ต่อเนื่องตัวที่ k สำหรับคลัสเตอร์ j

J คือจำนวนคลัสเตอร์

k คือข้อมูลตัวที่ k

l คือลำดับที่ของค่าที่เป็นไปได้ของตัวแปรที่ไม่ต่อเนื่อง

5. การหาระยะทางแบบ Euclidean (Euclidean Distance)

ระยะทางแบบ Euclidean ใช้ในการหาค่าความใกล้เคียงระหว่างจุดสองจุด และเมื่อนำมาประยุกต์ใช้ในกรณีของข้อมูลที่มีคุณลักษณะต่างๆ แต่ละจุดจึงเปรียบเสมือนค่าของแต่ละคุณลักษณะข้อมูล เพื่อใช้ในการคำนวณหาความใกล้เคียงกันของข้อมูล 2 ชิ้น ค่าระยะห่างที่คำนวณได้มีค่าน้อย แสดงว่าข้อมูลนั้นมีความคล้ายคลึงกันมากกว่าค่าระยะห่างที่มีค่ามากกว่า (Tan *et al.*, 2006)

$$Distance(x, y) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (17)$$

โดยที่ x_1 คือค่าคุณลักษณะ x ของข้อมูลที่ 1

x_2 คือค่าคุณลักษณะ x ของข้อมูลที่ 2

y_1 คือค่าคุณลักษณะ y ของข้อมูลที่ 1

y_2 คือค่าคุณลักษณะ y ของข้อมูลที่ 2

6. การคำนวณค่าน้ำหนักของคุณลักษณะข้อมูล

อัลกอริทึมการเลือกคุณลักษณะมี 2 ประเภท

1. วิธี Wrapper เป็นวิธีที่ใช้อัลกอริทึมการเรียนรู้เพื่อการจำแนกและประเมินผลให้กับคุณลักษณะข้อมูล เช่น ต้นไม้ตัดสินใจ และตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน
2. วิธีการกรอง (Filter Method) เป็นวิธีประเมินค่าคุณลักษณะข้อมูลให้สอดคล้องกับลักษณะพื้นฐานของข้อมูล เช่น อัลกอริทึมรีลีฟฟ์

สำหรับงานที่เกี่ยวข้องกับข้อมูลขนาดใหญ่ วิธีการกรองเป็นวิธีที่เหมาะสมมากกว่าวิธี Wrapper เนื่องจากใช้เวลาน้อยในการทำงาน (Hall, 2000)

ตัวอย่างของอัลกอริทึมการเลือกคุณลักษณะโดยวิธีการกรอง มีดังนี้

6.1 วิธีต้นไม้ตัดสินใจของ Hall

วิธีนี้คิดค้นโดย Hall (2007) เป็นวิธีการคำนวณค่าน้ำหนักคุณลักษณะจากต้นไม้ตัดสินใจ โดยพิจารณาจากค่าระดับที่น้อยที่สุดที่คุณลักษณะนั้นปรากฏบนต้นไม้ตัดสินใจ หากไม่พบคุณลักษณะนั้นบนต้นไม้ตัดสินใจแล้วค่าน้ำหนักสำหรับคุณลักษณะนั้นจะมีค่าเป็นศูนย์ โดยค่าน้ำหนักคุณลักษณะที่ได้นี้จะนำไปใช้ในตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก

โดยในการทดลองมี 5 กรณีในการเปลี่ยนแปลงขนาดของข้อมูลที่สุ่มเพื่อใช้ในการเรียนรู้ต้นไม้ตัดสินใจ สำหรับทุกกรณีนั้น การสร้างต้นไม้ตัดสินใจใช้วิธีการแบ็คกิ้ง และต้นไม้ตัดสินใจเป็นแบบไม่มีการตกแต่ง (Unpruned) แต่ละกรณีมีรายละเอียด ดังนี้

กรณีที่ 1 สร้างต้นไม้ตัดสินใจ 10 ต้น โดยที่แต่ละต้นเกิดจากการเรียนรู้ข้อมูลซึ่งได้จากการสุ่มข้อมูลแบบแทนที่ จำนวน 25% ของข้อมูลการเรียนรู้ทั้งหมด

กรณีที่ 2 สร้างต้นไม้ตัดสินใจ 10 ต้น โดยที่แต่ละต้นเกิดจากการเรียนรู้ข้อมูลซึ่งได้จากการสุ่มข้อมูลแบบแทนที่ จำนวน 50% ของข้อมูลการเรียนรู้ทั้งหมด

กรณีที่ 3 สร้างต้นไม้ตัดสินใจ 10 ต้น โดยที่แต่ละต้นเกิดจากการเรียนรู้ข้อมูลซึ่งได้จากการสุ่มข้อมูลแบบแทนที่ จำนวน 75% ของข้อมูลการเรียนรู้ทั้งหมด

กรณีที่ 4 สร้างต้นไม้ตัดสินใจ 10 ต้น โดยที่แต่ละต้นเกิดจากการเรียนรู้ข้อมูลซึ่งได้จากการสุ่มข้อมูลแบบแทนที่ จำนวน 100% ของข้อมูลการเรียนรู้ทั้งหมด

กรณีที่ 5 สร้างต้นไม้ตัดสินใจ 1 ต้น โดยที่ต้นไม้เกิดจากการเรียนรู้ข้อมูลการเรียนรู้ทั้งหมด

เมื่อได้ต้นไม้ตัดสินใจครบตามที่กำหนดแล้ว จึงคำนวณค่าน้ำหนักแต่ละคุณลักษณะจากต้นไม้ตัดสินใจในแต่ละต้น เมื่อคำนวณเรียบร้อยจึงหาค่าเฉลี่ยของค่าน้ำหนักแต่ละคุณลักษณะ อัลกอริทึมการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูลจากต้นไม้ตัดสินใจ แสดงดังภาพที่ 9

Algorithm Calculate attribute weights from decision tree of Hall (HW)

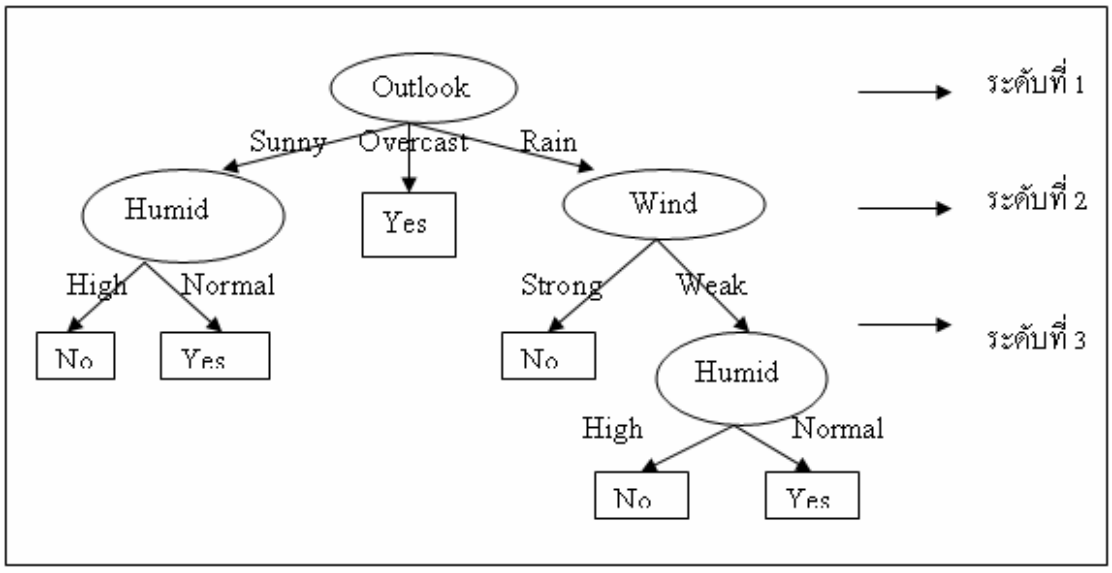
1. Repeat i times :
2. Randomly sample (with replacement) j% of the training data.
3. Learn an unpruned decision tree from the resampled data.
4. FOR each attribute in the training data DO :
5. IF the attribute is NOT tested in the tree THEN
6. Record a weight of 0.
7. ELSE
8. Let d be the minimum depth that the attribute is tested at.
9. Record a weight of $1/\sqrt{d}$
10. FOR each attribute in the training data DO :
11. Set the final weight equal to the average of the i weights.
12. Optionally remove from the data all attributes with zero weight.
13. Learn a naive Bayes model using the final attribute weights.

ภาพที่ 9 อัลกอริทึมการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูลจากต้นไม้ตัดสินใจ

ที่มา : Hall, 2007

ขั้นตอนวิธีการของอัลกอริทึมการคำนวณค่าน้ำหนักจากต้นไม้ตัดสินใจของ Hall

1. ในแต่ละรอบ $i = 1$ ถึง $i = k$
2. สุ่มข้อมูลแบบแทนที่จำนวน j% ของข้อมูลทั้งหมด
3. นำข้อมูลที่ได้จากข้อ 2. ไปเรียนรู้ด้วยการสร้างต้นไม้ตัดสินใจแบบไม่ตกแต่งค่า (unpruned)
4. คำนวณค่าน้ำหนักคุณลักษณะทุกคุณลักษณะ
   4.1. ถ้าไม่พบคุณลักษณะข้อมูลนั้นในต้นไม้ตัดสินใจบันทึกค่าน้ำหนักคุณลักษณะเป็นศูนย์
   4.2. ถ้าพบคุณลักษณะข้อมูลนั้นในต้นไม้ตัดสินใจบันทึกค่าน้ำหนักคุณลักษณะเป็นส่วนกลับของรากที่สองของระดับที่น้อยที่สุดที่พบคุณลักษณะนั้นในต้นไม้ตัดสินใจ
5. เมื่อดำเนินการครบ k รอบ จึงนำค่าน้ำหนักแต่ละคุณลักษณะทั้งหมด k ค่ามาเฉลี่ยและบันทึกเป็นค่าน้ำหนักของแต่ละคุณลักษณะ
6. นำค่าน้ำหนักที่ได้ในข้อ 5. ไปแทนค่าในการคำนวณของตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก



ภาพที่ 10 แสดงตัวอย่างต้นไม้ตัดสินใจสำหรับพิจารณาค่าน้ำหนักของคุณลักษณะข้อมูล
ที่มา : Mitchell (1997)

การคำนวณหาค่าน้ำหนักของแต่ละคุณลักษณะข้อมูล มีสมการดังนี้

$$w_{คุณลักษณะ} = \frac{1}{\sqrt{d}} \tag{18}$$

โดยที่ d แทนค่าระดับที่น้อยที่สุดของคุณลักษณะที่ปรากฏบนต้นไม้ตัดสินใจ

จากภาพที่ 10 ข้อมูลมีคุณลักษณะ 3 อย่าง คือ Outlook, Wind, Humid

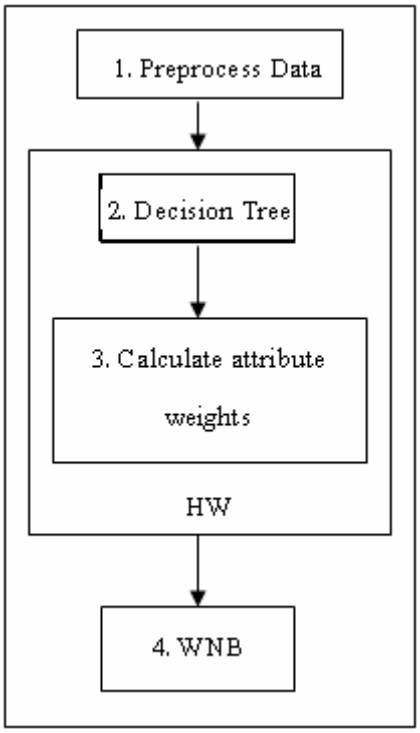
คุณลักษณะ Outlook อยู่ที่ระดับที่ 1

$$w_{Outlook} = \frac{1}{\sqrt{1}} = 1$$

คุณลักษณะ Humid อยู่ที่ระดับที่ 2 และ 3 เลือกระดับที่น้อยที่สุด นั่นคือระดับที่ 2

$$w_{Humid} = \frac{1}{\sqrt{2}} = 0.707$$

สำหรับลำดับการทำงานของงานวิจัยของ Hall สามารถอธิบายได้จากภาพที่ 11 ดังนี้



ภาพที่ 11 ลำดับการทำงานงานวิจัยของ Hall (2007)

ขั้นตอนที่ 1 เป็นการเตรียมข้อมูล คือแปลงคุณลักษณะข้อมูลสำหรับข้อมูลที่มีค่าต่อเนื่องไปสู่ค่าที่ไม่ต่อเนื่องเพื่อใช้เป็นข้อมูลนำเข้าสำหรับสร้างต้นไม้ตัดสินใจ

ขั้นตอนที่ 2 เป็นกระบวนการสร้างต้นไม้ตัดสินใจจากวิธีการแบ็คกิ้ง ซึ่งจะทำให้ได้ต้นไม้ตัดสินใจหลายต้นตามกำหนดไว้ โดยใช้ค่าพารามิเตอร์จาก 5 กรณี ที่ได้กล่าวไว้ข้างต้น คือ ข้อมูลสุ่มที่นำมาใช้ในการเรียนรู้ต้นไม้ตัดสินใจในแต่ละต้น โดยใช้วิธีการสุ่มแบบแทนที่ เช่น สำหรับกรณีที่ 1 สุ่มข้อมูลมา 25% จากข้อมูลการเรียนรู้ทั้งหมดคือจำนวน 900 ข้อมูล ในการสร้างต้นไม้ตัดสินใจ 10 ต้นนั้น ในแต่ละต้นจะทำการสุ่มข้อมูล 225 ข้อมูลมาแล้วนำข้อมูลนี้ไปสร้างต้นไม้ตัดสินใจ เมื่อสร้างต้นไม้ตัดสินใจเสร็จ จึงทำการสุ่มข้อมูลจากข้อมูลเดิม 900 ข้อมูล เพื่อสร้างต้นไม้ต้นต่อไป ดำเนินการจนเสร็จทั้ง 10 ต้น

ขั้นตอนที่ 3 การคำนวณค่าน้ำหนักคุณลักษณะจากต้นไม้ตัดสินใจ โดยค่าน้ำหนักแต่ละคุณลักษณะสามารถคำนวณได้จากค่าส่วนกลับของรากที่สองของระดับชั้นที่น้อยที่สุดที่คุณลักษณะนั้นปรากฏบนต้นไม้ตัดสินใจ แต่ถ้าไม่มีคุณลักษณะนั้นบนต้นไม้ตัดสินใจค่าน้ำหนักคุณลักษณะจะเท่ากับศูนย์ สำหรับกรณีที่ต้นไม้มีหลายต้นจะนำค่าน้ำหนักที่ได้ในแต่ละต้นไปหาค่าเฉลี่ยของค่า

น้ำหนักและกำหนดเป็นค่าน้ำหนักคุณลักษณะสำหรับการนำไปใช้ในตัวจำแนกไนอีฟ เบย์แบบน้ำหนักดังอัลกอริทึมในภาพที่ 9

ขั้นตอนที่ 4 การคำนวณค่าความน่าจะเป็นของข้อมูลทดสอบที่ให้ในแต่ละคลาสจากตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก น้ำหนักที่นำมาใช้ได้จากขั้นตอนที่ 3 สำหรับข้อมูลทดสอบได้จากการทำ 10 fold cross-validation ซึ่งเป็นวิธีการแบ่งข้อมูลเป็น 10 ส่วน และทำการทดลอง 10 การทดลอง โดยการทดลอง 1 ชุดจะนำข้อมูล 9 ส่วน สำหรับใช้ในขั้นตอนการเรียนรู้ (train data) และข้อมูลที่เหลือ 1 ส่วนใช้สำหรับการทดสอบ (test data)

6.2 อัลกอริทึมรีลีฟฟ์ (ReliefF Algorithm)

อัลกอริทึมรีลีฟฟ์คิดค้นโดย Kira และ Rendell ในปี 1992 เป็นอัลกอริทึมที่คำนวณค่าน้ำหนักคุณลักษณะข้อมูลจากการพิจารณาข้อมูลที่ใกล้เคียงกับข้อมูลสุ่ม (R) เช่น ข้อมูลมี 3 คลาส คือ คลาส A, B และ C เมื่อสุ่มข้อมูล R และพบว่า R อยู่คลาส A ข้อมูลใกล้เคียงจะมี 3 ข้อมูลคือ ข้อมูลที่ใกล้เคียงกับ R มากที่สุดของคลาส A แทนด้วย H และข้อมูลที่ใกล้เคียงกับ R มากที่สุดสำหรับคลาส B และ C แทนค่าด้วย M_B และ M_C

อัลกอริทึมรีลีฟฟ์ (Kira and Rendell, 1992)

Begin

$w[A] := 0.0$;

For i:=1 to m do

Randomly select an instance R

Find nearest hit H and nearest miss M for each class not class of R

For A:=1 to #all attribute do

$$w[A] := w[A] - \frac{diff(A,R,H)}{m} + \sum_{C \neq class(R)} \left(P(C) * \frac{diff(A,R,M_C)}{m} \right);$$

End.

ภาพที่ 12 แสดงอัลกอริทึมรีลีฟฟ์ในการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูล

ที่มา : Kononenko (1994)

จากภาพที่ 12 สามารถอธิบายตัวแปรได้ดังนี้

$$diff(A, ข้อมูลX, ข้อมูลY) = \frac{X(A) - Y(A)}{max(A) - min(A)} \quad (19)$$

โดยที่ m คือค่าจำนวนรอบของการคำนวณ ส่วนใหญ่ใช้ค่าจำนวนของข้อมูลเรียนรู้

R คือข้อมูลที่สุ่ม

H คือข้อมูลที่ใกล้กับข้อมูล R มากที่สุด และมีคลาสเดียวกับ R

M_C คือข้อมูลที่ใกล้กับข้อมูล R มากที่สุด และมีคลาส C ($C \neq Class(R)$)

$w[A]$ คือค่าน้ำหนักของคุณลักษณะ A

$X(A)$ และ $Y(A)$ คือค่าคุณลักษณะ A ของข้อมูล X และ Y ตามลำดับ

$P(C)$ คือค่าความน่าจะเป็นที่ข้อมูลที่จะให้คลาส C

6.3 อัตราส่วนเกน (Information Gain Ratio)

เป็นวิธีการคำนวณค่าน้ำหนักคุณลักษณะอีกวิธีหนึ่งจากต้นไม้ตัดสินใจ โดยค่าน้ำหนักคุณลักษณะสามารถคำนวณได้จากสมการที่ 20

$$w_i = \frac{GainRatio(A_i) \times m}{\sum_{i=1}^{m} GainRatio(A_i)} \quad (20)$$

โดยที่ w_i คือค่าน้ำหนักของคุณลักษณะที่ i

$GainRatio(A_i)$ คือค่าอัตราส่วนเกนของคุณลักษณะที่ i

m คือจำนวนคุณลักษณะ

A_i คือคุณลักษณะที่ i

การคำนวณน้ำหนักคุณลักษณะข้อมูลมีอีกมากมายสามารถอ่านเพิ่มเติมได้จาก Wettschereck *et al*. (1997) ซึ่งเป็นงานวิจัยที่แนะนำและเปรียบเทียบอัลกอริทึมการกำหนดค่าน้ำหนักของคุณลักษณะ

7. การวัดประสิทธิภาพ (Evaluation)

ในงานวิจัยนี้ใช้ค่าที่เกี่ยวข้องในการวัดประสิทธิภาพของการทำนายข้อมูลโดยตัวจำแนกไนอีฟ เบย์แบบน้ำหนักของคุณลักษณะข้อมูล (WNB) ดังนี้

7.1 ค่าความถูกต้องของการทำนาย (Accuracy)

ค่าความถูกต้องของการทำนาย (Accuracy) คือจำนวนข้อมูลทดสอบที่มีผลคำทำนายที่ถูกต้องตรงกับข้อมูลจริง ซึ่งจะคิดเป็นค่าร้อยละของจำนวนข้อมูลทดสอบที่มีคำทำนายที่ถูกต้องเปรียบเทียบกับข้อมูลทดสอบทั้งหมด

$$Accuracy = \frac{NA}{N} \times 100\% \tag{21}$$

โดยที่ NA คือจำนวนข้อมูลที่สอบที่ทำนายได้คลาสตรงตามข้อมูลจริง

N คือจำนวนข้อมูลทดสอบทั้งหมด

7.2 Root Relative Square Error (RRSE)

Root Relative Square Error (RRSE) (Witten and Frank, 2000) คือค่าที่ใช้อ้างความแตกต่างของความผิดพลาด (error) ซึ่งเป็นความสัมพันธ์ที่เกิดขึ้นจากการทำนายข้อมูล นั่นคือ เป็นค่าความแตกต่างระหว่างค่าความน่าจะเป็นของค่าที่แท้จริง (จะมี 2 ค่า คือ 0 และ 1) กับค่าความน่าจะเป็นที่ได้จากการทำนายโดยตัวจำแนก มีสมการดังนี้

$$\text{RRSE} = \sqrt{\frac{\sum_{k=1}^{N}\left(\sum_{i=1}^{C}\frac{(P_i - T_i)^2}{C}\right)_k}{\sum_{k=1}^{N}\left(\sum_{i=1}^{C}\frac{(T_i - T_p)^2}{C}\right)_k}} \tag{22}$$

โดยที่ P_i คือค่าความน่าจะเป็นที่คำนวณได้จากการทำนาย

T_i คือค่าความน่าจะเป็นที่แท้จริงของคำทำนาย มี 2 ค่า คือ 0 และ 1

T_p คือค่า Prior Proability สำหรับแต่ละคลาส

N คือจำนวนข้อมูลทดสอบทั้งหมด

C คือจำนวนคลาส

ในการพิจารณาค่า RRSE ประเมินผลเปรียบเทียบนั้น ค่า RRSE คือค่าความผิดพลาดที่จะเกิดขึ้นจากการทำนาย ซึ่งใช้เปรียบเทียบความแตกต่างของค่าความน่าจะเป็นที่คำนวณได้จริงกับค่าความน่าจะเป็นของคำตอบที่แท้จริง เพราะฉะนั้นวิธีทำนายที่มีค่า RRSE น้อยแสดงให้เห็นว่าวิธีนั้นมีประสิทธิภาพในการทำนายข้อมูลที่ดีกว่าวิธีที่มีค่า RRSE มากกว่า เนื่องจากผลความน่าจะเป็นของตัวทำนายแตกต่างจากความน่าจะเป็นที่แท้จริงน้อยกว่า

7.3 สถิติทดสอบ t-Test (pair t-Test)

สถิติทดสอบ t-Test เป็นการทดสอบสมมุติฐานเกี่ยวกับผลต่างของค่าเฉลี่ยข้อมูล 2 ชุด สำหรับข้อมูลสองชุดที่ไม่เป็นอิสระต่อกันนั้นคือ จะวัดค่าในเวลาที่ต่างกันหรือวัดค่าด้วยวิธีการวัดที่ต่างกัน เพื่อเปรียบเทียบความแตกต่างของประสิทธิภาพสำหรับสองอัลกอริทึมว่ามีความแตกต่างกันอย่างชัดเจนหรือไม่ วิธีการทดสอบคำนวณจากสมการดังนี้

1. กำหนดระดับนัยสำคัญ (α)
2. พิจารณาหาบริเวณวิกฤติ คือ $t > t_{\alpha;n\text{-}1}$ เมื่อ n แทนค่าจำนวนข้อมูลทั้งหมด
3. คำนวณค่า t

$$t = \frac{\overline{D}}{\sigma / \sqrt{n}} \tag{23}$$

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} \tag{24}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n - 1}} \tag{25}$$

โดยที่ D_i คือผลต่างของ 2 อัลกอริทึม ในชุดการทดลอง i

n คือจำนวนข้อมูลทั้งหมด

σ คือค่าของส่วนเบี่ยงเบนมาตรฐาน

เมื่อคำนวณค่า t ได้แล้ว พิจารณาว่าอยู่ในช่วงวิกฤติหรือไม่ ถ้าตกอยู่ในช่วงวิกฤติแสดงว่าอัลกอริทึมที่สองมีค่าของข้อมูลมากกว่าอัลกอริทึมที่หนึ่งอย่างมีนัยสำคัญ (สำหรับกรณีที่ D_i แทนค่าของข้อมูลที่ 2 – ข้อมูลที่ 1)

งานวิจัยที่เกี่ยวข้อง

1. งานวิจัยที่เกี่ยวข้องกับวิธีการคำนวณน้ำหนักคุณลักษณะ

งานวิจัยที่เกี่ยวข้องกับการปรับปรุงวิธีการคำนวณน้ำหนักคุณลักษณะข้อมูล มีรายละเอียดดังนี้

Cardie และ Howe (1997) คิดค้นวิธีการเลือกข้อมูลในขั้นตอนการสืบค้นข้อมูลแบบ Case-Based Reasoning โดยพิจารณาจากคุณลักษณะบนต้นไม้ตัดสินใจชนิด C4.5 แบบตกแต่ง (pruned) โดยนำคุณลักษณะข้อมูลที่เก็บไว้บนโหนดของต้นไม้ตัดสินใจใช้คำนวณค่าน้ำหนักจากค่าเกน (Information Gain) ผู้วิจัยรายนี้ใช้ค่าน้ำหนักแบบท้องถิ่น ซึ่งเป็นค่าน้ำหนักที่พิจารณาเฉพาะแต่ละชุดข้อมูล โดยที่การเลือกค่าน้ำหนักของคุณลักษณะข้อมูลที่เก็บไว้บนโหนดต้นไม้ตัดสินที่เหมาะสมกับข้อมูลทดสอบไปใช้กับอัลกอริทึม K Nearest Neighbor เพื่อค้นคืนข้อมูลภายใน Case base ที่ใกล้เคียงกับข้อมูลทดสอบ ข้อมูลที่ใช้ทดสอบมี 3 ชุดข้อมูลซึ่งนำมาจากกระบวนประมวลผลภาษาธรรมชาติ (Natural Language Processing-NLP) คือ p-o-s, sam class และ concept ผลจากการทดลองพบว่างานวิจัยนี้สามารถเพิ่มความแม่นยำของการทำนายคลาสได้อย่างมีนัยสำคัญเมื่อเปรียบเทียบกับอัลกอริทึมต้นไม้ตัดสินใจชนิด C4.5 และวิธีพื้นฐานการเลือกความถี่ของค่าคลาสที่มีมากที่สุด (Default) สำหรับ Kubat *et al*. (1993) ใช้วิธีการคำนวณค่าน้ำหนักคุณลักษณะจากค่าเกนเช่นเดียวกับ Cardie และ Howe (1997) แต่นำค่าน้ำหนักไปใช้กับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ผลที่ได้คือสามารถเพิ่มประสิทธิภาพให้แก่ตัวจำแนกไนอีฟ เบย์แบบมาตรฐานเช่นกัน

ต่อมา Ratanamahatana และ Gunopulos (2003) พบว่าประสิทธิภาพของต้นไม้ตัดสินใจ C4.5 ดีกว่าตัวจำแนกไนอีฟ เบย์ จึงนำวิธีทั้งสองทำงานร่วมกัน โดยนำเสนอตัวจำแนกการเลือกแบบเบย์เซียน (Selective Baysian classifier-SCB) ซึ่งจะเลือกใช้คุณลักษณะที่ปรากฏบนต้นไม้ตัดสินใจชนิด C4.5 เพียง 3 ระดับเท่านั้น ผลที่ได้คือวิธีของตัวจำแนก SCB มีประสิทธิภาพในการทำนายที่ดีกว่าตัวจำแนกไนอีฟ เบย์และต้นไม้ตัดสินใจชนิด C4.5 เพียงอย่างเดียว แต่ข้อด้อยของงานวิจัยนี้คือ ถ้าต้นไม้ตัดสินใจมีระดับชั้นที่มากกว่า 3 ระดับ จะไม่สามารถหาค่าน้ำหนักได้ถูกต้อง และ SCB ใช้วิธีการแบ็คกิ้งเพื่อสร้างต้นไม้ตัดสินใจหลายต้นโดยในแต่ละรอบใช้ข้อมูลประมาณ 10% ของข้อมูลการเรียนรู้ทั้งหมดเท่านั้น ซึ่งเป็นปริมาณที่น้อยสำหรับการคำนวณค่าคุณลักษณะข้อมูล และเมื่อนำค่าน้ำหนักมาใช้กับไนอีฟ เบย์แบบน้ำหนักจะให้ผลที่ด้อยกว่าอัลกอริทึมการคำนวณน้ำหนักอื่น เช่น อัลกอริทึมรีลีฟฟ์

Zhang และ Sheng (2004) ได้นำเสนอตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักและเปรียบเทียบวิธีการคำนวณค่าน้ำหนักคุณลักษณะโดยอัลกอริทึมหลากหลาย ได้แก่ วิธีอัตราส่วนเกน วิธี Hill Climbing วิธี Markov Chain Monte Carlo วิธี Hill Climbing ร่วมกับอัตราส่วนเกน และวิธี Markov Chain Monte Carlo ร่วมกับอัตราส่วนเกน โดยนำค่าน้ำหนักของคุณลักษณะที่ได้นั้นไปใช้ในตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ใช้ข้อมูลทดสอบจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) จำนวน 14 ชุด มีการประเมินผลโดยใช้ค่า AUC (Area under The ROC Curve) ผลที่ได้คือ ตัวจำแนกไนอีฟ เบย์ แบบค่าน้ำหนักทำงานดีกว่าตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน และพบว่าวิธีการคำนวณค่าน้ำหนักคุณลักษณะ Hill Climbing ร่วมกับอัตราส่วนเกนให้ผลลัพธ์ที่ดีที่สุดเมื่อนำมาใช้กับตัวจำแนกไนอีฟ เบย์ แบบใช้ค่าน้ำหนัก

สำหรับงานวิจัยที่เป็นแนวคิดหลักในการคำนวณน้ำหนักของคุณลักษณะโดยการจัดกลุ่มและต้นไม้ตัดสินใจของงานวิจัยฉบับนี้คือ วิธีการคำนวณค่าน้ำหนักของคุณลักษณะจากต้นไม้ตัดสินใจซึ่งคิดค้นโดย Hall (2007) เป็นวิธีการคำนวณค่าน้ำหนักของคุณลักษณะจากต้นไม้ตัดสินใจ โดยสามารถหาค่าน้ำหนักได้จากส่วนกลับของรากที่สองของระดับที่น้อยที่สุดที่คุณลักษณะนั้นปรากฏบนต้นไม้ตัดสินใจ ถ้าไม่พบคุณลักษณะนั้นบนต้นไม้ตัดสินใจ ค่าน้ำหนักสำหรับคุณลักษณะนั้นจะมีค่าเป็นศูนย์ โดยค่าน้ำหนักของคุณลักษณะที่ได้นี้จะนำไปใช้ในตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ในการศึกษาของ Hall ทำการทดลอง 2 แบบ คือ

การทดลองแบบที่ 1 คือการเปรียบเทียบประสิทธิภาพของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ Hall พัฒนาขึ้นจากต้นไม้ตัดสินใจกับตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน โดยในการทดลองมี 5 กรณี ดังรายละเอียดในหน้าที่ 23 และหน้าที่ 24 ในแต่ละกรณีมีการเปลี่ยนแปลงขนาดของข้อมูลที่สุ่มเพื่อใช้ในการเรียนรู้ต้นไม้ตัดสินใจ สำหรับทุกกรณีนั้น การสร้างต้นไม้ตัดสินใจใช้วิธีการแบ็คกิ้ง และต้นไม้ตัดสินใจเป็นแบบไม่มีการตกแต่ง (Unpruned) และการทดลองแบบที่ 2 คือการเปรียบเทียบประสิทธิภาพของอัลกอริทึมที่ใช้คำนวณน้ำหนักคุณลักษณะ ได้แก่ อัลกอริทึมของ Hall (2007) และอัลกอริทึมการคำนวณน้ำหนักอื่นๆ คือ อัตราส่วนเกน, อัลกอริทึมรีลีฟฟ์, วิธีการเลือกคุณลักษณะความสัมพันธ์, กฎการเลือกของเบย์, ตัวจำแนกการเลือกของเบย์เซียน และ NBTree

ข้อมูลของงานวิจัยของ Hall (2007) ได้มาจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) จำนวน 28 ชุดข้อมูล โดยที่ชุดข้อมูลนั้นมีจำนวนข้อมูลที่น้อยที่สุดคือ 57 ข้อมูล และจำนวนข้อมูลที่มากที่สุดคือ 5,000 ข้อมูล และข้อมูลมีทั้งคุณลักษณะที่เป็นค่าต่อเนื่องอย่างเดียว หรือค่าไม่ต่อเนื่องอย่างเดียว หรือมีทั้งค่าที่ไม่ต่อเนื่องและค่าที่ต่อเนื่องอยู่ร่วมกัน สำหรับการประเมินประสิทธิภาพพิจารณาจากค่า Root Relative Square Error (RRSE) และค่า Area under the ROC Curve (AUC) และคำนวณค่าความแตกต่างแบบมีนัยสำคัญ t-Test ที่ระดับ 5%

จากการเปรียบเทียบประสิทธิภาพพบว่า การทำงานของตัวจำแนกไนอีฟ เบย์ที่ใช้ค่าน้ำหนักจากต้นไม้ตัดสินใจมีประสิทธิภาพที่ดีกว่าอย่างมีนัยสำคัญโดยการทดสอบด้วย RRSE และค่า AUC เมื่อนำไปใช้กับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก พิจารณาผลการทดลองทั้ง 2 การทดลองดังนี้

การทดลองแบบที่หนึ่งคือ เมื่อพิจารณาเปรียบเทียบกับตัวจำแนกไนอีฟ เบย์มาตรฐานกับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีของ Hall ในกรณีที่กล่าวไว้ข้างต้นพบว่า ค่า RRSE สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีของ Hall มีค่าที่ดีกว่าตัวจำแนกไนอีฟ เบย์มาตรฐาน 12 ชุดข้อมูล และด้อยกว่า 2 ชุดข้อมูล สำหรับค่า AUC สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีของ Hall มีค่าที่ดีกว่าตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน 4 ชุดข้อมูล สำหรับค่า RRSE ในแต่ละกรณีพบว่ามีค่าใกล้เคียงกัน สำหรับการทดลองแบบที่สองคือ เมื่อพิจารณาเปรียบเทียบกับอัลกอริทึมคำนวณค่าน้ำหนักอื่นพบว่า ค่า RRSE สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีของ Hall มีค่าที่ดีกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากอัลกอริทึมการคำนวณน้ำหนักอื่น 13 ชุด

ข้อมูล และด้อยกว่า 1 ชุดข้อมูล สำหรับค่า AUC สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีของ Hall มีค่าที่ดีกว่าค่า AUC ของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากอัลกอริทึมการคำนวณน้ำหนักอื่น ยกเว้นข้อมูล KR ของอัลกอริทึม NBTree เท่านั้นที่มีค่า AUC ที่มากกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีของ Hall

สำหรับงานวิจัยของ Hall (2007) พบว่า ประสิทธิภาพการคำนวณค่าน้ำหนัก เมื่อพิจารณาขนาดชุดข้อมูลพบว่า เมื่อข้อมูลมีจำนวนน้อยคือน้อยกว่า 400 ข้อมูล จะมีค่า RRSE ที่มากกว่าตัวจำแนกไนอีฟ เบย์มาตรฐาน แสดงให้เห็นว่าค่า RRSE มีค่าที่มากกว่า นั่นคือมีประสิทธิภาพในการทำนายข้อมูลที่ด้อยกว่าตัวจำแนกไนอีฟ เบย์มาตรฐาน แต่เมื่อใช้วิธีการวัดทางสถิติพบว่าประสิทธิภาพการทำงานของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ได้จากวิธีต้นไม้ตัดสินใจของ Hall ไม่แตกต่างกับตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน ที่ระดับนัยสำคัญ 5%

2. งานวิจัยที่เกี่ยวข้องกับวิธีการจัดกลุ่มข้อมูล

การจัดกลุ่มด้วยอัลกอริทึมเค-มีนมีประสิทธิภาพการจัดกลุ่มที่ดี อัลกอริทึมตัวอย่างเช่น Forman และ Zhang (2000) ศึกษาและเปรียบเทียบประสิทธิภาพการแบ่งกลุ่มข้อมูลของเค-มีน, K-Harmonic Means และ Expectation-Maximization (EM) ผลการทดสอบพบว่าขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีน มีประสิทธิภาพในการจัดกลุ่มดีกว่าอัลกอริทึม K-Harmonic Means ประมาณ 6 เท่า และดีกว่าอัลกอริทึม Expectation-Maximization ประมาณ 200 เท่า ต่อมางานวิจัยของ Sinka และ Corne (2002) นำเสนอการจัดกลุ่มข้อมูลที่มีจำนวนข้อมูลมาก โดยใช้ขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน ทำการจัดกลุ่มข้อมูลที่เป็นเอกสารเว็บจำนวน 11 กลุ่ม กลุ่มละ 1000 เอกสาร แต่การทดลองจะเปรียบเทียบครั้งละ 2 กลุ่ม โดยเลือกทั้งกลุ่มที่มีความแตกต่างกันมาก เช่น กลุ่มของธุรกิจกับกีฬา และ กลุ่มที่มีความแตกต่างกันน้อยเช่น กลุ่มของกีฬาฟุตบอลกับกีฬาเทนนิส และทำการวัดผลโดยดูจากความถูกต้องของผลของการจัดกลุ่มเทียบกับผลเฉลย ซึ่งผลที่ได้พบว่าการจัดกลุ่มโดยขั้นตอนวิธีการจัดกลุ่มแบบเค-มีน ให้ผลของการจัดกลุ่มมีค่าความถูกต้องสูงสุดที่ประมาณ 90% แต่ค่าเฉลี่ยของผลการทดลองจัดกลุ่มทั้งหมดอยู่ที่ประมาณ 50%

งานของวิวัฒนะเจริญชัยและศรีวิหค (2003) นำเทคนิคการทำเหมืองข้อมูลไปใช้ในการแบ่งกลุ่มข้อมูลลูกค้า (Clustering) ที่ใช้งานอินเทอร์เน็ตแบงก์กิ้ง และทำการเปรียบเทียบอัลกอริทึมที่ใช้ในการจัดกลุ่มระหว่างขั้นตอนวิธีการจัดกลุ่มข้อมูลแบบเค-มีนและนิวรอลเน็ตเวิร์กอัลกอริทึม

มูลมาก ส่วนนิวรอลเน็ตเวิร์กอัลกอริทึมจะสามารถจัดกลุ่มได้ดีกว่าเมื่อข้อมูลมีจำนวนน้อย

การจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) และวิธีลำดับชั้นแบบ Agglomerative ซึ่งจากงานของ Norusis (2006) กล่าวถึงรายละเอียดของวิธีการจัดกลุ่มมี 3 แบบ คือ วิธีการลำดับชั้น วิธีเค-มีน และการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) และวิธีลำดับชั้นแบบAgglomerative ซึ่งกล่าวถึงความเหมาะสม ข้อดี และข้อเสียของแต่ละแบบ ในส่วนการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) และวิธีลำดับชั้นแบบ Agglomerative สามารถดำเนินการกับข้อมูลที่มีขนาดใหญ่ด้วยเวลาที่รวดเร็วและสามารถจัดการได้ทั้งตัวแปรที่เป็นค่าต่อเนื่องและไม่ต่อเนื่อง และสามารถคำนวณค่าจำนวนกลุ่มที่เหมาะสมได้จากค่า BIC (Schwarz's Baysian Criterion) (Fraley and Raftery, 1998)

งานวิจัยของ Zhang *et al.*(1996) คิดค้นวิธีการจัดกลุ่มที่ใช้วิธีลำดับชั้นที่ชื่อว่า BIRCH ซึ่งเหมาะสำหรับการจัดกลุ่มกับข้อมูลที่มีขนาดใหญ่ BIRCH สามารถจัดกลุ่มได้ดีด้วยการพิจารณาข้อมูลทีละตัว โดยในงานวิจัยนี้ได้เปรียบเทียบผลลัพธ์การจัดกลุ่มระหว่าง วิธี BIRCH และ วิธี CLARANS พบว่า BIRCH สามารถใช้จัดกลุ่มกับข้อมูลที่มีขนาดใหญ่ได้ดีกว่า CLARANS เมื่อพิจารณาจากคุณภาพและความเร็ว ซึ่งงานวิจัยนี้แสดงให้เห็นว่าต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) มีประสิทธิภาพที่ดีในการจัดกลุ่มข้อมูล และมีการนำวิธีของงานวิจัยนี้ไปประยุกต์ใช้กับการจัดกลุ่มแบบสองขั้นตอนของโปรแกรมสำเร็จรูปทางสถิติ

ตารางที่ 1 ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการคำนวณน้ำหนักคุณลักษณะ

งานวิจัย	เทคนิค	ข้อมูล	ผล
Cardie และ Howe (1997)	วิธีการค้นคืนข้อมูลบน Case-Based โดยใช้ค่าน้ำหนักจากค่าเกนของคุณลักษณะข้อมูลที่เก็บไว้บนโหนดต้นไม้ตัดสินใจชนิด C4.5 และเลือกคุณลักษณะบนต้นไม้ตัดสินใจให้เหมาะสมกับข้อมูลทดสอบ ก่อนนำค่าน้ำหนักไปใช้กับอัลกอริทึม k-Nearest Neighbor เพื่อค้นคืนข้อมูล	ข้อมูล 3 ชุดข้อมูลซึ่งนำมาจากกระบวนการภาษาธรรมชาติ (Natural Language Processing-NLP) คือ p-o-s, sam class และ concept	ค่าน้ำหนักจากค่าเกนที่ใช้กับ k-Nearest Neighbor เพิ่มความแม่นยำของการทำนายคลาสได้อย่างมีนัยสำคัญเมื่อเปรียบเทียบกับวิธีต้นไม้ตัดสินใจชนิด C4.5
Ratanamahatana และ Gunopulos (2003)	ใช้คุณลักษณะที่ปรากฏบนต้นไม้ตัดสินใจเพียง 3 ระดับบนในการคำนวณค่าน้ำหนักคุณลักษณะให้ตัวจำแนกการเลือกของเบย์เซียน	ข้อมูลจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) 10 ชุดข้อมูล	ตัวจำแนกการเลือกของเบย์เซียนมีความแม่นยำที่ดีกว่าอัลกอริทึม C4.5 และตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน
Zhang และ Sheng (2004)	คำนวณค่าน้ำหนักคุณลักษณะโดยวิธีอัตราส่วนเกน วิธี Hill Climbing วิธี Markov Chain Monte Carlo Hill Climbing ร่วมกับอัตราส่วนเกน และวิธี Markov Chain Monte Carlo ร่วมกับอัตราส่วนเกนโดยนำไปใช้ในตัวจำแนกไนอีฟ เบย์ แบบค่าน้ำหนัก	ข้อมูลจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) จำนวน 14 ชุด	การคำนวณค่าน้ำหนักโดยวิธี Hill Climbing ร่วมกับอัตราส่วนเกนให้ผลลัพธ์ที่ดีที่สุดเมื่อใช้กับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก

ตารางที่ 1 (ต่อ)

งานวิจัย	เทคนิค	ข้อมูล	ผล
Mark Hall (2007)	คำนวณค่าน้ำหนักคุณลักษณะจากต้นไม้ตัดสินใจ และเปรียบเทียบกับวิธีอื่นคือ อัตราส่วนเกน อัลกอริทึมรีลีฟฟ์ วิธีการเลือกคุณลักษณะความสัมพันธ์ กฎการเลือกของเบย์ ตัวจำแนกการเลือกของเบย์เซี่ยน และ NBTree โดยนำไปใช้ในตัวจำแนกไนอีฟ เบย์ แบบค่าน้ำหนัก	ข้อมูลจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) จำนวน 28 ชุด	ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งคำนวณจากต้นไม้ตัดสินใจ (NBH) ของข้อมูลทดสอบ 12 ชุด มีประสิทธิภาพที่ดีกว่าตัวจำแนกไนอีฟเบย์มาตรฐานอย่างมีนัยสำคัญ และตัวจำแนก NBH ของข้อมูลทดสอบ 13 ชุด มีประสิทธิภาพที่ดีกว่าวิธีการคำนวณน้ำหนักอื่นอย่างมีนัยสำคัญ

ตารางที่ 2 ตารางสรุปผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่ม

งานวิจัย	เทคนิค	ข้อมูล	ผล
Forman และ Zhang (2000)	ศึกษาเปรียบเทียบวิธีการแบ่งกลุ่มข้อมูลเค-มีน, K-Harmonic Means และ Expectation-Maximization	ข้อมูลทางด้านเครือข่าย (Network)	การแบ่งกลุ่มแบบเค-มีนมีประสิทธิภาพในการแบ่งกลุ่มดีกว่า K-Harmonic Means ประมาณ 6 เท่า และดีกว่าอัลกอริทึม EM ประมาณ 200 เท่า
Sinka และ Corne (2002)	การจัดกลุ่มข้อมูลที่มีจำนวนข้อมูลมาก โดยใช้วิธีเค-มีน ทำการจัดกลุ่มข้อมูลที่เป็นเอกสารเว็บทดลอง เปรียบเทียบครั้งละ 2 กลุ่ม	ชุดข้อมูลเป็นเอกสารเว็บจำนวน 11 กลุ่ม กลุ่มละ 1000 เอกสาร	การจัดกลุ่มโดยวิธีเค-มีน ให้ค่าความถูกต้องสูงสุดประมาณ 90% โดยค่าเฉลี่ยของผลการจัดกลุ่มทั้งหมดอยู่ที่ประมาณ 50%
Fraley และ Raftery (1998)	ใช้วิธีลำดับชั้นแบบ Agglomerative ในการกำหนดค่าเริ่มต้นให้กับอัลกอริทึม EM ส่วนของจำนวนกลุ่มใช้ค่า Bayesian Information Criterion (BIC) เพื่อหาจำนวนกลุ่มที่เหมาะสม	ข้อมูลของการตรวจโรคเบาหวานจากการสังเกต 145 ข้อมูล มีตัวแปร 3 ตัว คือ กลูโคส, อินซูลิน และระดับของการต้านอินซูลิน	วิธีลำดับชั้นแบบ Agglomerative ร่วมกับอัลกอริทึม EM สามารถเพิ่มประสิทธิภาพการจัดกลุ่มได้ดีกว่ากระบวนการจัดกลุ่มอัลกอริทึม EM และ อัลกอริทึมเค-มีน

ตารางที่ 2 (ต่อ)

งานวิจัย	เทคนิค	ข้อมูล	ผล
Zhang et al. (1996)	วิธีการจัดกลุ่มที่ใช้วิธีลำดับชั้นแบบ BIRCH	ข้อมูลภาพจริง 100 รูป ขนาด 512x1024 พิกเซล ที่มี 2 ความยาวคลื่น คือ NIR (Near Infrared band) และ VIS (Visible wavelength band)	วิธีการจัดกลุ่มแบบ BIRCH สามารถใช้จัดกลุ่มกับข้อมูลที่มีขนาดใหญ่ได้ดีกว่า CLARANS เมื่อพิจารณาจากคุณภาพและความเร็ว

อุปกรณ์และวิธีการ

อุปกรณ์

1. ฮาร์ดแวร์

คอมพิวเตอร์ตั้งโต๊ะ หน่วยประมวลผล AMD Athlon64 X2 Dual Core 3600+ ความเร็ว 1.90 กิกะเฮิร์ต หน่วยความจำหลัก DDR2 667 ขนาด 2048 เมกะไบต์ หน่วยความจำสำรอง 160 กิกะไบต์ แบบ SATA II

2. ซอฟแวร์

1) Java Development Kits (JDK) 5.0 Update 10 พร้อมกับ NetBeans 5.5
2) โปรแกรม Weka version 3.5.5
3) โปรแกรมสำเร็จรูปทางสถิติ
4) ระบบปฏิบัติการ Microsoft Windows XP Service Pack 2

วิธีการ

1. ข้อมูลสำหรับทำการทดลอง (Data set)

ข้อมูลที่ใช้ในการทดลองสำหรับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) นำมาจากข้อมูลของมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) ซึ่งเป็นชุดข้อมูลมาตรฐานที่ใช้ทดสอบในงานด้านการทำเหมืองข้อมูล ซึ่งในงานวิจัยนี้ใช้ 4 ชุดข้อมูล คือ Chess End-Game (KR), Balance Scale Weight& Distance Database (Bal_sc), Wave Form และ German Credit dataset (Crd) รายละเอียดข้อมูลดังตารางที่ 4 สำหรับการเลือกข้อมูล 4 ชุดข้อมูลนี้เนื่องจากต้องการพิจารณาในเรื่องขนาดข้อมูลใหญ่และเล็ก คือ จำนวนมากที่สุด 5,000 ข้อมูลและจำนวนน้อยที่สุดคือ 625 ข้อมูล และในเรื่องของคุณลักษณะข้อมูลจากจำนวนคุณลักษณะและประเภทของ

ตารางที่ 3 ตารางแสดงรายละเอียดของชุดข้อมูลที่ใช้ในงานวิจัยนี้

Data set	Instance	Numeric	Nominal	Class
Bal_sc	625	4	0	3
Crd	1,000	7	13	2
KR	3,196	0	36	2
Wave	5,000	40	0	3

ตัวอย่างของชุดข้อมูล Balance Scale Weight& Distance Database (Bal_sc) ซึ่งมี 625 ข้อมูล มีคุณลักษณะที่มีค่าต่อเนื่อง 4 คุณลักษณะ และสามารถแบ่งได้ 3 คลาส แสดงดังตาราง 4

ตารางที่ 4 แสดงตัวอย่างค่าคุณลักษณะที่ 1 ถึงคุณลักษณะที่ 5 (V1-V4) และคลาสของข้อมูล Bal_sc

V1	V2	V3	V4	Class
1	1	1	1	B
1	1	1	2	R
1	1	1	3	R
1	1	1	4	R
1	1	1	5	R
1	1	2	1	R
1	1	2	2	R

ตัวอย่างของชุดข้อมูล German Credit (Crd) ซึ่งมี 1,000 ข้อมูล มีคุณลักษณะที่มีค่าต่อเนื่อง 7 คุณลักษณะ คุณลักษณะที่ไม่ต่อเนื่อง 13 คุณลักษณะ และสามารถแบ่งได้ 2 คลาส ดังตารางที่ 5

ตารางที่ 5 แสดงตัวอย่างค่าคุณลักษณะที่ 1 ถึงคุณลักษณะที่ 7 (V1-V7) และคลาสของข้อมูล German Credit dataset (Crd)

V1	V2	V3	V4	V5	V6	V7	Class
6	A34	A43	1169	A65	A75	4	Yes
48	A32	A43	5951	A61	A73	2	No
12	A34	A46	2096	A61	A74	2	Yes
42	A32	A42	7882	A61	A74	2	Yes
24	A33	A40	4870	A61	A73	3	No
36	A32	A46	9055	A65	A73	2	Yes
24	A32	A42	2835	A63	A75	3	Yes

ตัวอย่างของชุดข้อมูล Chess End-Game (KR) ซึ่งมี 3,196 ข้อมูล มีคุณลักษณะที่มีค่าไม่ต่อเนื่อง 36 คุณลักษณะ และสามารถแบ่งได้ 2 คลาส แสดงดังตาราง 6

ตารางที่ 6 แสดงตัวอย่างค่าคุณลักษณะที่ 1 ถึงค่าคุณลักษณะที่ 7 (V1-V7) และคลาสของข้อมูล KR

V1	V2	V3	V4	V5	V6	V7	Class
f	f	f	f	f	t	t	won
f	f	f	f	f	f	t	nowin
f	f	f	f	f	t	t	won
f	f	f	f	t	t	t	nowin
f	f	f	f	t	t	t	won

เมื่อ f และ t คือค่าคุณลักษณะที่เป็นไปได้ของคุณลักษณะที่ 1 ถึงคุณลักษณะที่ 7 (V1-V7)

ตัวอย่างของชุดข้อมูล Wave Form ซึ่งมี 5,000 ข้อมูล มีคุณลักษณะที่มีค่าต่อเนื่อง 40 คุณลักษณะ และสามารถแบ่งได้ 3 คลาส แสดงดังตาราง 7

ตารางที่ 7 แสดงตัวอย่างค่าคุณลักษณะที่ 1 ถึงคุณลักษณะที่ 7 (V1-V7) และคลาสของข้อมูล Wave

V1	V2	V3	V4	V5	V6	V7	Class
-1.23	-1.56	-1.75	-0.28	0.6	2.22	-1.23	0
-0.69	2.43	0.61	2.08	2.3	3.25	-0.69	1
-0.12	-0.94	1.29	2.59	2.42	3.55	-0.12	1
0.86	0.29	2.19	-0.02	1.13	2.51	0.86	2
1.16	0.37	0.4	-0.59	2.66	1	1.16	0
0	0.77	1.32	0.29	-1.28	0.84	0	2
0.87	1.07	-0.65	1.46	0.84	2.7	0.87	2

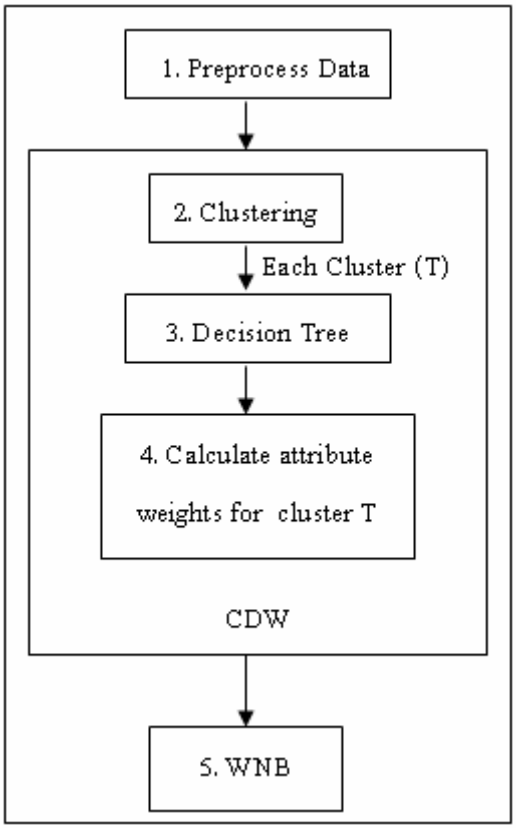
2. ขั้นตอนการทดลอง

ในงานวิจัยนี้จึงได้นำเสนอวิธีการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูลจากการจัดกลุ่มและต้นไม้ตัดสินใจ ซึ่งจะปรับปรุงเพิ่มเติมจากงานวิจัยของ Hall ซึ่งมีการคำนวณค่าน้ำหนักของคุณลักษณะข้อมูลจากต้นไม้ตัดสินใจ ค่าน้ำหนักของแต่ละคุณลักษณะจะมีค่าเดียวกันเมื่อนำไปใช้กับทุกข้อมูลทดสอบ ในงานวิจัยนี้ต้องการพิสูจน์ว่า เมื่อมีการจัดกลุ่มข้อมูลการเรียนรู้ก่อน แล้วนำแต่ละกลุ่มข้อมูลที่ได้มาสร้างต้นไม้ตัดสินใจเพื่อคำนวณหาค่าน้ำหนักแต่ละคุณลักษณะ ในส่วนของการทดสอบข้อมูล พิจารณาว่า ข้อมูลทดสอบมีความใกล้เคียงกับกลุ่มข้อมูลใดมากที่สุดแล้วเลือกใช้ค่าน้ำหนักของแต่ละคุณลักษณะของข้อมูลกลุ่มนั้นไปใช้ในตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) วิธีการนี้สามารถเพิ่มประสิทธิภาพและความแม่นยำในการทำนายของตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก สำหรับวิธีการจัดกลุ่มที่นำมาใช้จะเลือกทดลอง 2 แบบ คือ

1) การจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative เนื่องจากสามารถจัดกลุ่มข้อมูลที่มีขนาดใหญ่ได้ดี และจัดการกับข้อมูลที่มีค่าแบบต่อเนื่องและไม่ต่อเนื่องได้ และงานวิจัยที่สามารถแสดงว่าการจัดกลุ่มแบบสองขั้นตอนนี้มีประสิทธิภาพที่ดี เช่น งานวิจัยของ Norusis (Norusis et al.,2002) สำหรับงานวิจัยของ

2) การจัดกลุ่มโดยใช้อัลกอริทึมเค-มีน มีประสิทธิภาพการจัดกลุ่มที่ดีและใช้เวลาน้อย งานวิจัยจำนวนมากที่นำอัลกอริทึมเค-มีนไปประยุกต์ใช้ในการจัดกลุ่ม เช่น งานวิจัยของ Sinka และ Corne (2002) งานวิจัยของวิวัฒนะเจริญชัยและศรีวิหค (2003) และมีการนำเค-มีน ไปเปรียบเทียบกับการจัดกลุ่มวิธีการอื่นๆ และพบว่าเค-มีน ให้ประสิทธิภาพในการจัดกลุ่มที่ดี

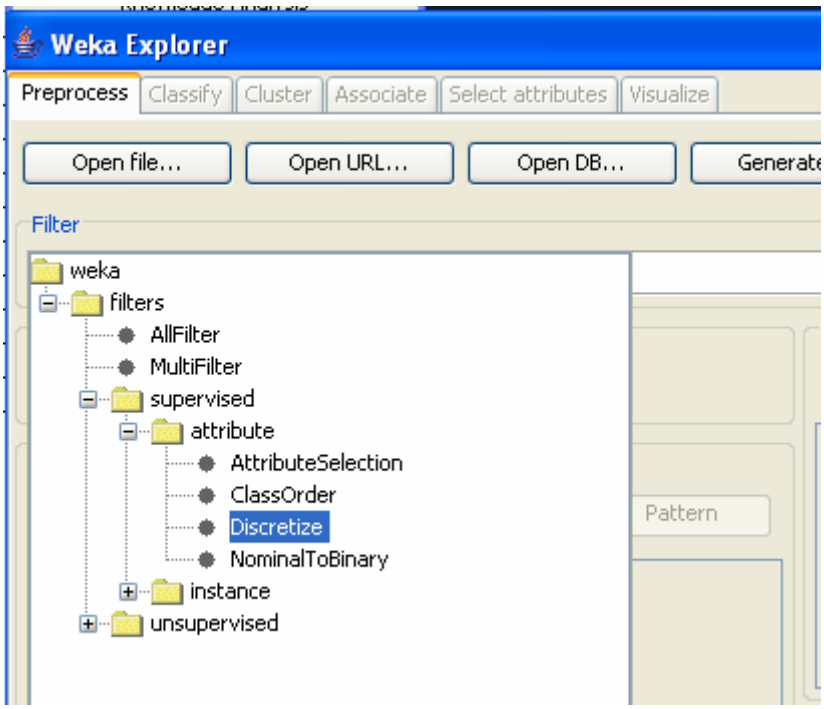
ในงานวิจัยนี้จะมีการทดลองเพื่อใช้ในการเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึม 3 อัลกอริทึม คือ ตัวจำแนกไนอีฟ เบย์ มาตรฐาน (NB) ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักคุณลักษณะจากต้นไม้ตัดสิน ของ Hall (HW) และ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักคุณลักษณะจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) ลำดับขั้นตอนการทำงานของการศึกษาครั้งนี้แสดงในภาพที่ 13



ภาพที่ 13 แสดงลำดับการทำงานของอัลกอริทึมคำนวณค่าน้ำหนักคุณลักษณะด้วยวิธีการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW)

1. การเตรียมข้อมูล

การเตรียมข้อมูลสำหรับการทดลอง เนื่องจากงานวิจัยฉบับนี้ใช้อัลกอริทึมที่ทำงานได้เหมาะสมกับข้อมูลที่มีค่าไม่ต่อเนื่อง สำหรับข้อมูลที่มีค่าต่อเนื่องจึงจำเป็นต้องแปลงข้อมูลเป็นค่าไม่ต่อเนื่องก่อน โดยใช้อัลกอริทึม Discretization (สำหรับแปลงข้อมูลที่มีค่าต่อเนื่องไปสู่ค่าที่ไม่ต่อเนื่อง) ของโปรแกรม Weka 3.5.5



ภาพที่ 14 ตัวอย่างโปรแกรม Weka ในส่วนของการแปลงข้อมูลให้เป็นค่าที่ไม่ต่อเนื่อง (Discretize)

ภาพที่ 14 แสดงลำดับการเลือกวิธีการแปลงข้อมูล (Discretize) บนหน้าจอของโปรแกรม Weka สำหรับกรณีที่ต้องแปลงข้อมูลจากค่าที่ต่อเนื่องไปสู่ค่าที่ไม่ต่อเนื่อง จากภาพโดยเลือกที่ Filter คุณลักษณะข้อมูล (attribute) แบบ supervised และเลือก Discretize

V1	V2	V3	V4	V5
-0.81	1.59	-0.69	1.16	4.22
0.59	0.77	-0.61	1	1.8
-0.15	0.13	2.27	2.39	4
-0.3	-0.42	0.25	-0.61	-1.39
-1.45	2.71	3.04	3.21	4.26
0.28	0.97	-1.01	-2.34	-1.89
-1.09	-0.44	1.15	0.17	2.1
0.5	-1.23	-0.09	0.31	2.22
-0.23	-0.44	1.04	0.38	0.53
-1.2	-1.04	-0.06	1.24	-1.41
-0.53	0.8	0.1	0.25	0.59
-0.17	-0.27	-0.46	2.09	-1.32
0.58	1.69	-0.77	0.36	0.37

ภาพที่ 15 ตัวอย่างค่าคุณลักษณะที่ 1 ถึงคุณลักษณะที่ 5 (V1-V5) ของข้อมูล Wave ก่อนการแปลงข้อมูลในโปรแกรม Weka

V1	V2	V3	V4	V5
\'(0.515-inf)\"	\'(-0.695-0.525]\"	\'(0.395-1.315]\"	\'(2.825-inf)\"	\'(4.075-inf)\"
\'(0.515-inf)\"	\'(-0.695-0.525]\"	\'(0.395-1.315]\"	\'(1.465-2.225]\"	\'(1.935-2.735]\"
\'(-0.435-0.515]\"	\'(2.075-inf)\"	\'(2.165-inf)\"	\'(2.825-inf)\"	\'(4.075-inf)\"
\'(-0.435-0.515]\"	\'(-0.695-0.525]\"	\'(-0.705-0.395]\"	\'(-inf--0.135]\"	\'(-inf-0.195]\"
\'(0.515-inf)\"	\'(2.075-inf)\"	\'(2.165-inf)\"	\'(2.825-inf)\"	\'(4.075-inf)\"
\'(0.515-inf)\"	\'(-inf--0.695]\"	\'(-inf--0.705]\"	\'(-inf--0.135]\"	\'(0.195-0.885]\"
\'(-inf--0.435]\"	\'(0.525-1.165]\"	\'(-0.705-0.395]\"	\'(1.465-2.225]\"	\'(2.735-4.075]\"
\'(-inf--0.435]\"	\'(-0.695-0.525]\"	\'(-0.705-0.395]\"	\'(1.465-2.225]\"	\'(-inf-0.195]\"
\'(-inf--0.435]\"	\'(0.525-1.165]\"	\'(-0.705-0.395]\"	\'(-0.135-0.745]\"	\'(-inf-0.195]\"
\'(-inf--0.435]\"	\'(-0.695-0.525]\"	\'(0.395-1.315]\"	\'(-inf--0.135]\"	\'(0.195-0.885]\"
\'(0.515-inf)\"	\'(-0.695-0.525]\"	\'(-0.705-0.395]\"	\'(-0.135-0.745]\"	\'(-inf-0.195]\"
\'(-0.435-0.515]\"	\'(-0.695-0.525]\"	\'(1.315-2.165]\"	\'(-inf--0.135]\"	\'(0.195-0.885]\"
\'(0.515-inf)\"	\'(-inf--0.695]\"	\'(-0.705-0.395]\"	\'(-0.135-0.745]\"	\'(0.195-0.885]\"

ภาพที่ 16 ตัวอย่างค่าคุณลักษณะที่ 1 ถึงคุณลักษณะที่ 5 (V1-V5) ของข้อมูล Wave Form หลังการแปลงข้อมูลให้เป็นค่าไม่ต่อเนื่องในโปรแกรม Weka

จากภาพที่ 16 ข้อมูลแต่ละคุณลักษณะจะมีค่าเป็นช่วง ดังตัวอย่าง เช่น คุณลักษณะ V1 มีค่าที่เป็นไปได้ คือ '\'(-inf--0.435]\", '\'(-0.435-0.515]\", '\'(0.515-inf)\" และคุณลักษณะ V2 มีค่าที่เป็นไปได้ คือ '\'(-inf--0.695]\", '\'(-0.695-0.525]\", '\'(0.525-1.165]\", '\'(1.165-2.075]\" และ '\'(2.075-inf)\" ต่อมาจึงเปลี่ยนค่าที่ได้จากภาพที่ 16 โดยเปรียบเทียบค่าคุณลักษณะกับค่าลำดับช่วงข้อมูล จึงได้ข้อมูลดังภาพที่ 17 นั่นคือจากค่าคุณลักษณะ V1 มี 3 ค่าลำดับช่วง คือ 0, 1, 2 และคุณลักษณะ V2 มี 5 ค่าลำดับช่วง คือ 0, 1, 2, 3, 4

V1	V2	V3	V4	V5
2	1	2	5	5
2	1	2	3	3
1	4	4	5	5
1	1	1	0	0
2	4	4	5	5
2	0	0	0	1
0	2	1	3	4
0	1	1	3	0
0	2	1	1	0
0	1	2	0	1
2	1	1	1	0
1	1	3	0	1
2	0	1	1	1

ภาพที่ 17 ตัวอย่างการแปลงค่าคุณลักษณะจากภาพที่ 16 ให้เป็นค่าลำดับช่วงข้อมูล

2. การจัดกลุ่ม

ในการศึกษาครั้งนี้มีการเปรียบเทียบ 2 วิธีการจัดกลุ่มข้อมูลก่อนที่จะนำไปสร้างต้นไม้ตัดสินใจของแต่ละกลุ่มข้อมูล นั่นคือ

1) การจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative โดยใช้จากโปรแกรมการจัดกลุ่มแบบสองขั้นตอน (Two Step Clustering) จากโปรแกรมสำเร็จรูปทางสถิติ

2) การจัดกลุ่มแบบเค-มีน

โดยจะกำหนดจำนวนกลุ่ม (K) มีค่าตั้งแต่ 2 กลุ่มถึง 6 กลุ่ม และจำนวนกลุ่มที่คัดเลือกได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion) (Fraley และ Raftery, 1998) เมื่อจัดกลุ่มข้อมูลเรียบร้อยจึงเข้าสู่ขั้นตอนที่ 3

3. การสร้างต้นไม้ตัดสินใจโดยใช้วิธีการแบ็คกิ้ง

ขั้นตอนนี้เป็นกระบวนการสร้างต้นไม้ตัดสินใจให้แก่ข้อมูลในแต่ละกลุ่มที่ได้รับการจัดกลุ่มในข้อ 2. ซึ่งในการทดลองนี้จะใช้ต้นไม้ตัดสินใจแบบไม่มีการตัดแต่งค่า (Unpruned) โดยใช้อัลกอริทึม J48 และมีพารามิเตอร์ 2 ตัว คือ ค่า i แทนจำนวนต้นไม้ตัดสินใจที่สร้างขึ้น และค่า j คือร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ โดยในการทดลองมีการเปรียบเทียบค่าพารามิเตอร์ 4 กรณี มีค่าดังนี้

กรณีที่ 1 สร้างต้นไม้ตัดสินใจจำนวน 10 ต้น (i=10) ซึ่งแต่ละต้นได้มาจากการเรียนรู้ข้อมูลจำนวน 25% จากข้อมูลทั้งหมด (j=25%)

กรณีที่ 2 สร้างต้นไม้ตัดสินใจจำนวน 10 ต้น (i=10) ซึ่งแต่ละต้นได้มาจากการเรียนรู้ข้อมูลจำนวน 50% จากข้อมูลทั้งหมด (j=50%)

กรณีที่ 3 สร้างต้นไม้ตัดสินใจจำนวน 10 ต้น (i=10) ซึ่งแต่ละต้นได้มาจากการเรียนรู้ข้อมูลจำนวน 75% จากข้อมูลทั้งหมด (j=75%)

กรณีที่ 4 สร้างต้นไม้ตัดสินใจจำนวน 1 ต้น (i=1) ซึ่งต้นไม้ได้มาจากการเรียนรู้ข้อมูลทั้งหมด (j=100%)

สำหรับในขั้นตอนนี้เป็นขั้นตอนสำหรับการสร้างต้นไม้ตัดสินใจเพื่อคำนวณหาค่าน้ำหนักให้ในแต่ละกลุ่มข้อมูลที่ได้รับการจัดกลุ่มในข้อ 2. ยกตัวอย่างเช่น ข้อมูลการเรียนรู้มีทั้งหมด 1,000 ข้อมูล นำข้อมูลทั้งหมดนี้ไปจัดกลุ่มตามกำหนด สมมุติจัดกลุ่ม 5 กลุ่ม หลังจากการจัดกลุ่ม 5 กลุ่มพบว่า กลุ่มที่ 1 มี 200 ข้อมูล, กลุ่มที่ 2 มี 300 ข้อมูล, กลุ่มที่ 3 มี 150 ข้อมูล, กลุ่มที่ 4 มี 175 ข้อมูล และกลุ่มที่ 5 มี 75 ข้อมูล จึงนำข้อมูลในแต่ละกลุ่มเข้าสู่ข้อที่ 3. คือการสร้างต้นไม้ตัดสินใจให้แต่ละกลุ่ม

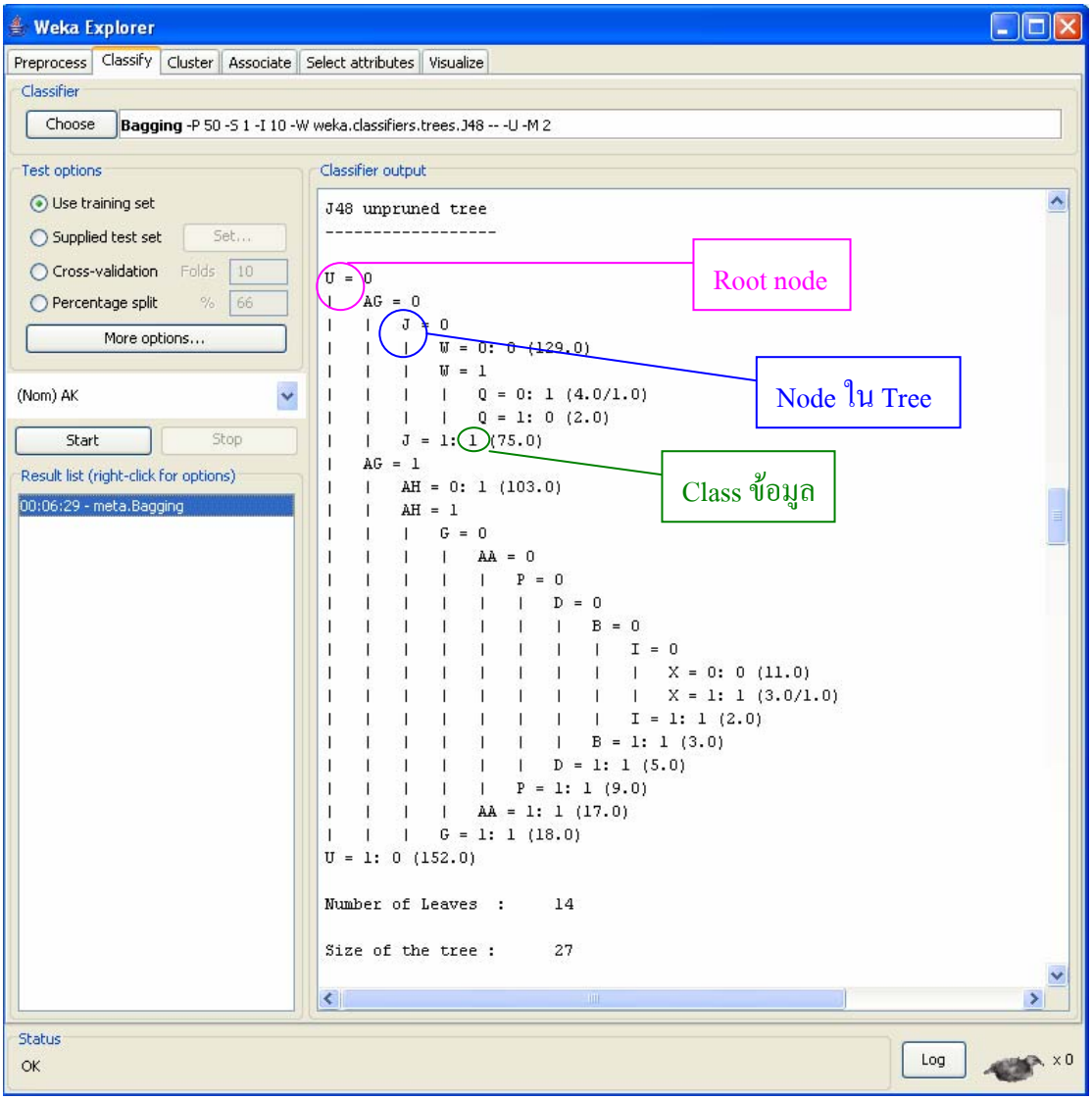
เริ่มพิจารณาที่กลุ่มที่ 1 มีข้อมูล 200 ข้อมูล ทำการทดลองในกรณีที่ 1 คือสร้างต้นไม้ 10 ต้น จากข้อมูลการเรียนรู้ 25% การทำงานเริ่มจากการสร้างต้นไม้ต้นที่ 1 โดยเริ่มสุ่มข้อมูล 50 ข้อมูลจากจำนวนทั้งหมด 200 ข้อมูล และนำ 50 ข้อมูลนี้มาสร้างต้นไม้ตัดสินใจ ต่อมาการสร้างต้นที่ 2 โดยเริ่มสุ่มข้อมูล 50 ข้อมูลจากข้อมูลชุดเดิมจำนวน 200 ข้อมูล และนำ 50 ข้อมูลนี้มาสร้างต้นไม้ตัดสินใจต้นที่ 2 ทำซ้ำเช่นนี้จนกระทั่งได้ต้นไม้ครบ 10 ต้น สำหรับเงื่อนไขการสร้างต้นไม้ในกรณีที่ 2-4 นั้นใช้หลักการเดียวกัน

การสร้างต้นไม้ให้กับกลุ่มที่ 2 ถึงกลุ่มที่ 5 นั่นมีขั้นตอนการทำงานเช่นเดียวกับกลุ่มที่ 1 โดยทุกกลุ่มจะมีการสร้างต้นไม้ตัดสินใจ 4 กรณี

4. คำนวณค่าน้ำหนักของคุณลักษณะข้อมูลจากต้นไม้ตัดสินใจ

สำหรับในแต่ละกลุ่มข้อมูลจะมีกลุ่มต้นไม้ตัดสินใจของกลุ่ม นั่นคือ นำกลุ่มต้นไม้ตัดสินใจเหล่านั้นมาคำนวณค่าน้ำหนักคุณลักษณะให้แต่ละกลุ่มข้อมูล ดังนี้

1. พิจารณาต้นไม้ตัดสินใจทีละต้น และบันทึกค่าน้ำหนักของแต่ละคุณลักษณะจากค่าส่วนกลับของรากที่สองของระดับที่น้อยที่สุดที่คุณลักษณะนั้นปรากฏบนต้นไม้ตัดสินใจ พิจารณาได้จากภาพที่ 18



ภาพที่ 18 ตัวอย่างต้นไม้ตัดสินใจโดยใช้วิธีการแบ็คกิ้งของอัลกอริทึม J48 (C4.5) ในโปรแกรม Weka

จากภาพที่ 18 สามารถอธิบายได้ว่าโหนด U คือรากของต้นไม้ เพราะอยู่ในระดับชั้นที่ 1 ของต้นไม้ตัดสินใจ และเมื่อ U มีค่า 0 คุณลักษณะต่อมาคือ AG ซึ่งจะอยู่ในลำดับชั้นที่ 2 ของต้นไม้ตัดสินใจ ถ้าโหนด AG มีค่าเป็น 0 ก็พิจารณาจากโหนด J ถ้าโหนด J เป็น 1 คลาสข้อมูลมีค่าเป็นคลาส 1 ซึ่งค่าคลาสของข้อมูลพิจารณาจากตัวเลขหลังเครื่องหมายโคลอน (:) แต่ถ้าโหนด J มีค่าเป็น 0 ไปพิจารณาที่โหนด W ถ้าโหนด W มีค่า 0 คลาสข้อมูลมีค่าเป็นคลาส 1 แต่ถ้าโหนด W มีค่า 1 ไปพิจารณาที่โหนด Q ซึ่งเป็นโหนดสุดท้ายที่ใช้พิจารณาคลาสข้อมูล คือถ้าโหนด Q มีค่าเป็น 0

การคำนวณหาค่าน้ำหนักของแต่ละคุณลักษณะข้อมูล มีดังนี้

$$w_{\text{คุณลักษณะ}} = \frac{1}{\sqrt{d}}$$

โดยที่ d แทนค่าระดับที่น้อยที่สุดของคุณลักษณะที่ปรากฏบนต้นไม้ตัดสินใจ ถ้าคุณลักษณะไม่ปรากฏบนต้นไม้ตัดสินกำหนดให้ค่าน้ำหนักมีค่าเป็นศูนย์

ตัวอย่างการคำนวณน้ำหนักคุณลักษณะข้อมูล (Node) จากภาพที่ 18

โหนด U อยู่ที่ระดับที่ 1

$$w_U = \frac{1}{\sqrt{1}} = 1$$

คุณลักษณะ X อยู่ที่ระดับที่ 10

$$w_X = \frac{1}{\sqrt{10}} = 0.316$$

2. เมื่อคำนวณค่าน้ำหนักคุณลักษณะครบทุกต้นแล้วจึงหาค่าเฉลี่ยของค่าน้ำหนักสำหรับแต่ละคุณลักษณะจากต้นไม้ตัดสินใจทั้งหมด ผลที่ได้คือค่าน้ำหนักของแต่ละคุณลักษณะข้อมูลเพื่อนำไปใช้กับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก

5. การทำนายข้อมูลของตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB)

การทดสอบประสิทธิภาพการทำงานของตัวจำแนกใช้การแบ่งข้อมูลทดสอบโดยวิธี 10 fold cross-validation กับทุกชุดข้อมูล นั่นคือ สำหรับข้อมูลทั้งหมดในชุดข้อมูลนำมาแบ่งออกเป็น 10 ส่วน ซึ่งสามารถสร้างการทดลองได้ 10 การทดลองโดยในแต่ละการทดลองใช้ข้อมูล 9 ส่วนเป็นข้อมูลการเรียนรู้ (training data) และข้อมูล 1 ส่วนที่เหลือเป็นข้อมูลการทดสอบ

สำหรับการเลือกใช้ค่าน้ำหนักคุณลักษณะกับข้อมูลทดสอบ พิจารณาจากค่าความแตกต่างที่น้อยที่สุดระหว่างค่าคุณลักษณะของข้อมูลทดสอบกับค่าคุณลักษณะเฉลี่ยของแต่ละกลุ่มข้อมูลโดยใช้วิธีการคำนวณค่าระยะทางแบบ Euclidean และจึงนำค่าน้ำหนักคุณลักษณะของกลุ่มที่มีระยะห่างกับข้อมูลทดสอบที่น้อยที่สุดนั้นไปแทนในสมการคำนวณความน่าจะเป็นของตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) และคำนวณหาค่าความน่าจะเป็นที่ข้อมูลทดสอบจะให้คลาสในแต่ละคลาสแล้วจึงนำไปประเมินผลด้วยค่าความถูกต้องของการทำนาย (Accuracy) และ Root Relative Square Error (RRSE) ต่อไป

ผลและวิจารณ์

ผล

ผลการทดลองจากการทำนายของตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB)

ในการทดสอบการทำงานของโมเดลนั้นใช้ชุดข้อมูล 4 ชุด ดังนี้

1. ข้อมูล Balance Scale Weight& Distance Database (Bal_sc)

เป็นชุดข้อมูลขนาดเล็ก คือมีจำนวน 625 ข้อมูล และมี 4 คุณลักษณะ แต่ละคุณลักษณะเป็นค่าที่ต่อเนื่อง สามารถแบ่งได้เป็น 3 คลาส คือ Balance, Right และ Left เมื่อนำมาทดสอบกับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) NB และ NBH ได้ผลดังตารางที่ 8 – 13 ดังนี้

ตารางที่ 8 แสดงค่าความถูกต้องของการทำนายข้อมูล Bal_sc สำหรับ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	85.77%	**90.79%**	88.41%	86.72%	86.34%	83.33%
2	10	50%	85.07%	**90.95%**	88.57%	88.57%	86.66%	84.60%
3	10	75%	89.97%	**91.90%**	90.95%	90.00%	89.36%	87.61%
4	1	100%	86.78%	**91.58%**	88.57%	86.82%	85.39%	82.85%

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 8 พบว่า ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากวิธีการ NBCT จะให้ค่าความถูกต้องของการทำนายข้อมูลที่น้อยลงเมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดที่จำนวนกลุ่ม 2 กลุ่ม

ตารางที่ 9 แสดงค่าความถูกต้องของการทำนายข้อมูล Bal_sc สำหรับ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	88.59%	91.26%	**91.74%**	91.11%	83.49%	86.50%
2	10	50%	90.00%	90.63%	**90.79%**	89.52%	89.04%	88.88%
3	10	75%	90.52%	**91.74%**	91.58%	91.11%	88.41%	86.66%
4	1	100%	89.53%	91.58%	**91.74%**	89.36%	83.01%	80.47%

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 9 พบว่า ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากวิธีการ NBCK จะให้ค่าความถูกต้องของการทำนายข้อมูลที่น้อยลง เมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดอยู่ที่จำนวนกลุ่ม 2 กลุ่ม สำหรับกรณีที่ 3 และจำนวน 3 กลุ่ม สำหรับกรณีที่ 1, 2 และ 4

ตารางที่ 10 แสดงการเปรียบเทียบค่าความถูกต้องของการทำนายข้อมูล Bal_sc ระหว่าง ตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) ซึ่งได้มาจากวิธีต้นไม้ตัดสินใจของ Hall (HW) และวิธีการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	**91.74%**	**91.74%**	90.79% ๐	**91.74%**
2	10	50%	**91.74%**	91.42%	90.95%	90.79%
3	10	75%	**91.74%**	91.42%	91.90%	**91.74%**
4	1	100%	**91.74%**	**91.74%**	91.58%	**91.74%**

โดยที่ ๐ แสดงว่าผลนั้นมีค่าความถูกต้องของการทำนายน้อยกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 10 จะทำการคัดเลือกค่าความถูกต้องของการทำนาย สำหรับแต่ละวิธีการจัดกลุ่มข้อมูลที่มีค่ามากที่สุดของข้อมูล Bal_sc ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 8 และตารางที่ 9 พบว่าการจัดกลุ่มทั้ง 2 วิธีให้ค่าที่ใกล้เคียงกันไม่แตกต่างกันนัก และมีค่าน้อยกว่าค่าความถูกต้องของการทำนายของตัวจำแนก NB และตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี Hall (HW) แต่ไม่แตกต่างกันอย่างมีนัยสำคัญ 5%

ข้อมูล Bal_sc สำหรับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	66.72 ± 5.12	**62.67 ± 5.09**	65.96 ± 7.01	66.90 ± 8.12	66.57 ± 8.27	71.36 ± 7.47
2	10	50%	65.33 ± 9.57	**58.25 ± 4.91**	61.63 ± 7.60	61.21 ± 5.96	63.61 ± 7.98	67.43 ± 7.79
3	10	75%	63.55 ± 3.95	**61.87 ± 4.08**	62.60 ± 4.12	63.17 ± 5.10	63.86 ± 5.01	66.50 ± 6.58
4	1	100%	66.72 ± 8.10	**62.08 ± 3.93**	65.99 ± 7.25	66.41 ± 8.29	68.42 ± 9.77	71.28 ± 8.07

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 11 พบว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการ NBCT มีค่า RRSE ที่เพิ่มขึ้นเมื่อจำนวนกลุ่มเพิ่มขึ้นและมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 2 กลุ่ม

ตารางที่ 12 แสดงค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล Bal_sc สำหรับ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ได้มาจากการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	63.87 ± 4.97	62.01 ± 4.49	**61.86 ± 3.66**	62.94 ± 4.84	70.75 ± 5.67	66.63 ± 5.39
2	10	50%	59.79 ± 5.76	**58.68 ± 5.67**	58.84 ± 3.87	60.30 ± 5.58	60.78 ± 4.59	61.73 ± 5.35
3	10	75%	62.39 ± 3.87	61.87 ± 4.31	**61.81 ± 3.77**	62.09 ± 4.38	66.46 ± 6.11	65.62 ± 5.70
4	1	100%	65.33 ± 5.02	61.92 ± 4.27	**61.89 ± 3.71**	64.18 ± 4.58	70.54 ± 5.98	71.41 ± 7.02

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 12 พบว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการ NBCK มีค่า RRSE ที่เพิ่มขึ้นเมื่อจำนวนกลุ่มเพิ่มขึ้นและมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 2 กลุ่ม สำหรับกรณีที่ 2 และจำนวน 3 กลุ่ม สำหรับกรณีที่ 1, 3 และ 4

ตารางที่ 13 แสดงการเปรียบเทียบค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล Bal_sc ระหว่าง ตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) จากต้นไม้ตัดสินใจของ Hall (HW) และ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	65.02 ± 3.59	**57.98 ± 4.55**	62.67 ± 5.09 * -	61.86 ± 3.66 * -
2	10	50%	65.02 ± 3.59	**58.08 ± 4.83**	58.25 ± 4.91 *	58.68 ± 5.67 *
3	10	75%	65.02 ± 3.59	**58.13 ± 3.87**	61.87 ± 4.08 * -	61.81 ± 3.77 * -
4	1	100%	65.02 ± 3.59	63.99 ± 3.73	62.08 ± 3.93 * +	**61.89 ± 3.71 * +**

โดยที่ * แสดงว่าผลนั้นมีค่า RRSE น้อยกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5%

\+ แสดงว่าผลนั้นมีค่า RRSE น้อยกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งได้มาจากวิธีต้นไม้ตัดสินใจของ Hall (HW) อย่างมีนัยสำคัญ 5%

\- แสดงว่าผลนั้นมีค่า RRSE มากกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5% ซึ่งได้มาจาก Hall (HW) อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 13 จะทำการคัดเลือกค่า RRSE สำหรับแต่ละวิธีการจัดกลุ่มข้อมูลที่ให้ค่าน้อยที่สุดของข้อมูล Bal_sc ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 11 และตารางที่ 12 พบว่า ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจาก NBCK ให้ค่า RRSE ที่ดีกว่าวิธี NBCT และมีค่าที่ดีกว่าตัวจำแนก NB อย่างมีนัยสำคัญ แต่มีค่า RRSE ที่ด้อยกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี Hall (HW)

สรุปผลการทดลองจากตารางที่ 8 ถึงตารางที่ 13 สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) ของข้อมูล Bal_sc จะมีค่าความถูกต้องของการทำนายใกล้เคียงตัวจำแนกไนอีฟ เบย์ มาตรฐาน (NB) และตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) แต่สำหรับค่า Root Relative Square Error (RRSE) จะให้ค่าที่ดีกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5% แต่ด้อยกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW)

เมื่อพิจารณาวิธีการจัดกลุ่มที่เหมาะสมกับข้อมูล Bal_sc พบว่าวิธีของอัลกอริทึมเค-มีน จะมีค่าความถูกต้องของการทำนาย และค่า RRSE ที่ดีกว่าวิธีการจัดกลุ่มสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative

สำหรับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) ของข้อมูล Bal_sc จะมีค่าความถูกต้องของการทำนาย และค่า RRSE ที่ดี เมื่อใช้กำหนดจำนวนกลุ่ม 2 กลุ่มสำหรับการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative (NBCT) และ จำนวนกลุ่ม 3 กลุ่ม สำหรับวิธีการจัดกลุ่มแบบเค-มีน (NBCK)

2. ข้อมูล German Credit dataset (Crd)

เป็นชุดข้อมูลที่มีขนาดเล็ก คือมีจำนวน 1,000 ข้อมูล และมี 7 คุณลักษณะที่เป็นค่าต่อเนื่อง และ 13 คุณลักษณะที่เป็นค่าไม่ต่อเนื่อง สามารถแบ่งได้เป็น 2 คลาส คือ Yes และ No เมื่อนำมาทดสอบกับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ได้ผลดังตารางที่ 14 – 19 ดังนี้

ตารางที่ 14 แสดงค่าความถูกต้องของการทำนายข้อมูล Crd สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	74.40%	**74.40%**	73.60%	72.00%	70.90%	70.90%
2	10	50%	76.00%	**76.50%**	74.80%	74.10%	72.70%	73.00%
3	10	75%	75.10%	**75.10%**	74.70%	74.60%	73.30%	73.60%
4	1	100%	74.50%	**74.50%**	74.40%	74.00%	73.80%	74.00%

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 14 พบว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT) จะให้ค่าความถูกต้องของการทำนายข้อมูลน้อยลงเมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดที่จำนวนกลุ่ม 2 กลุ่ม

ตารางที่ 15 แสดงค่าความถูกต้องของการทำนายข้อมูล Crd สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	73.90%	**74.00%**	73.10%	72.10%	71.80%	71.60%
2	10	50%	74.10%	**74.60%**	73.90%	73.70%	74.20%	73.40%
3	10	75%	74.60%	75.00%	73.70%	73.60%	**75.30%**	74.00%
4	1	100%	74.20%	**74.80%**	73.10%	74.00%	74.20%	73.90%

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 15 พบว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการ NBCK จะให้ค่าความถูกต้องของการทำนายข้อมูลน้อยลงเมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดที่จำนวนกลุ่ม 2 กลุ่ม สำหรับกรณีที่ 1, 2 และ 4 และจำนวน 5 กลุ่ม สำหรับกรณีที่ 3

ตารางที่ 16 แสดงการเปรียบเทียบความถูกต้องของการทำนายข้อมูล Crd ระหว่าง ตัวจำแนกไนอีฟเบย์มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) และตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	74.60%	**75.30%**	74.40%	74.00%
2	10	50%	74.60%	74.90%	**76.50%** +	74.60%
3	10	75%	74.60%	75.00%	75.10%	**75.30%**
4	1	100%	74.60%	**75.10%**	74.50%	74.80%

โดยที่ + แสดงว่าผลนั้นมีค่าความถูกต้องของการทำนายข้อมูลมากกว่าตัวจำแนกไนอีฟเบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 16 จะทำการคัดเลือกค่าความถูกต้องของการทำนาย สำหรับแต่ละวิธีการจัดกลุ่มข้อมูลที่มีค่ามากที่สุดของข้อมูล Crd ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 14 และ ตารางที่ 15 พบว่าการจัดกลุ่มทั้ง 2 วิธีให้ค่าที่ใกล้เคียงกันแต่ไม่แตกต่างกันนัก และส่วนใหญ่มีค่าน้อยกว่าค่าความถูกต้องของการทำนายของตัวจำแนก NB และตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี Hall (HW) แต่ไม่แตกต่างกันอย่างมีนัยสำคัญ 5%

ตารางที่ 17 แสดงค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล Crd สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ได้มาจากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	89.82 ± 3.23	**89.18 ± 3.09**	89.87 ± 2.30	91.24 ± 1.54	92.39 ± 2.63	92.77 ± 2.46
2	10	50%	89.30 ± 3.39	**88.88 ± 3.52**	89.42 ± 2.85	89.49 ± 2.93	90.77 ± 3.23	90.59 ± 2.49
3	10	75%	89.28 ± 3.72	**89.00 ± 3.92**	89.18 ± 3.25	89.22 ± 3.21	89.98 ± 3.29	89.84 ± 2.86
4	1	100%	89.89 ± 4.48	**89.33 ± 4.23**	89.89 ± 3.55	89.84 ± 4.08	90.22 ± 3.07	90.01 ± 2.97

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 17 พบว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจาก NBCT มีค่า RRSE ที่เพิ่มขึ้นเมื่อจำนวนกลุ่มเพิ่มขึ้นและมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 2 กลุ่ม

ตารางที่ 18 แสดงค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล Crd สำหรับ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ได้มาจากการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	89.59 ± 2.94	**89.46 ± 3.49**	90.18 ± 2.83	90.98 ± 2.56	91.30 ± 2.12	91.71 ± 1.90
2	10	50%	89.48 ± 3.39	**89.16 ± 4.20**	89.75 ± 3.08	89.38 ± 2.94	89.99 ± 2.79	90.27 ± 2.85
3	10	75%	89.29 ± 3.57	**89.27 ± 4.16**	89.66 ± 3.49	89.28 ± 3.61	89.21 ± 3.28	89.27 ± 2.89
4	1	100%	**89.18 ± 4.22**	89.20 ± 4.42	89.94 ± 4.27	90.41 ± 3.45	89.62 ± 3.40	89.73 ± 3.01

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 18 พบว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจาก NBCK มีค่า RRSE ที่เพิ่มขึ้นเมื่อจำนวนกลุ่มเพิ่มขึ้นและมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 2 กลุ่ม สำหรับกรณีที่ 1, 2 และ 3 และเมื่อใช้จำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion) สำหรับกรณีที่ 4

ตารางที่ 19 แสดงการเปรียบเทียบค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล Crd ระหว่างตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) และ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	90.35 ± 5.22	89.22 ± 3.85	**89.18 ± 3.09**	89.46 ± 3.49 *
2	10	50%	90.35 ± 5.22	89.13 ± 4.34	**88.88 ± 3.52** *	89.16 ± 4.20 *
3	10	75%	90.35 ± 5.22	89.17 ± 4.57	**89.00 ± 3.92** *	89.27 ± 4.16 *
4	1	100%	90.35 ± 5.22	**89.15 ± 4.65**	89.33 ± 4.23 *	89.18 ± 4.42 *

โดยที่ * แสดงว่าผลนั้นมีค่า RRSE น้อยกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 19 จะทำการคัดเลือกค่า RRSE สำหรับแต่ละวิธีการจัดกลุ่มข้อมูลที่มีค่าน้อยที่สุดของข้อมูล Crd ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 17 และตารางที่ 18 พบว่าการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative จะมีค่า RRSE ที่ดีกว่าวิธีของเค-มีน รวมทั้งมีค่า RRSE ที่ดีกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5% และมีค่า RRSE ที่ดีตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี Hall (HW) แต่ไม่แตกต่างกันอย่างมีนัยสำคัญ 5%

สรุปผลการทดลองจากตารางที่ 14 ถึงตารางที่ 19 ซึ่งเป็นการเปรียบเทียบค่าความถูกต้องของการทำนายและ ค่า Root Relative Square Error (RRSE) ของข้อมูล Crd ระหว่างตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) และ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) พบว่า สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งได้มาจากวิธี CDW ของข้อมูล Crd จะมีค่าความถูกต้องของการทำนายใกล้เคียงตัวจำแนก NB และตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งได้มาจากวิธี Hall (HW) แต่สำหรับค่า RRSE จะให้ค่าที่ดีกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5% แต่มีค่า RRSE ที่มากกว่าเมื่อเทียบกับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งได้มาจากวิธี Hall (HW)

เมื่อพิจารณาวิธีการจัดกลุ่มที่เหมาะสมกับข้อมูล Crd พบว่าวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative จะมีค่าความถูกต้องของการทำนาย และค่า RRSE ที่ดีกว่าวิธีการจัดกลุ่มแบบเค-มีน

สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี CDW ของข้อมูล Crd จะมีค่าความถูกต้องของการทำนาย และค่า RRSE ที่ดี เมื่อกำหนดจำนวนกลุ่ม 2 กลุ่ม ให้แก่การจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative และการจัดกลุ่มแบบเค-มีน

3. ข้อมูล Chess End-Game (KR)

เป็นข้อมูลขนาดใหญ่ มีจำนวน 3,196 ข้อมูล และมี 36 คุณลักษณะ แต่ละคุณลักษณะเป็นค่าที่ไม่ต่อเนื่อง สามารถแบ่งได้เป็น 2 คลาส คือ won และ nowon เมื่อนำมาทดสอบกับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ได้ผลดังตารางที่ 20 – 25 ดังนี้

ตารางที่ 20 แสดงค่าความถูกต้องของการทำนายข้อมูล KR สำหรับ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ได้จากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	92.06%	92.15%	92.25%	93.00%	**93.37%**	92.71%
2	10	50%	92.06%	92.50%	92.09%	93.12%	**93.31%**	93.18%
3	10	75%	92.09%	91.59%	92.43%	92.71%	92.90%	**93.09%**
4	1	100%	92.46%	92.71%	92.12%	93.21%	93.06%	**93.53%**

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 20 พบว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการ NBCT จะให้ค่าความถูกต้องของการทำนายข้อมูลที่ดีเมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดอยู่ที่จำนวนกลุ่ม 5 กลุ่ม สำหรับกรณีที่ 1 และ 2 และจำนวน 6 กลุ่ม สำหรับกรณีที่ 3 และ 4

ตารางที่ 21 แสดงค่าความถูกต้องของการทำนายข้อมูล KR สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ได้มาจากการจัดกลุ่มแบบเค-มีนและต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	92.09%	92.00%	92.31%	92.15%	**93.34%**	93.15%
2	10	50%	92.09%	91.34%	92.18%	92.84%	93.18%	**93.37%**
3	10	75%	91.71%	91.06%	92.06%	92.59%	93.28%	**93.65%**
4	1	100%	92.50%	91.96%	92.87%	92.71%	92.53%	**93.18%**

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 21 พบว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจาก NBCK จะให้ค่าความถูกต้องของการทำนายข้อมูลที่ดีเมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดอยู่ที่จำนวนกลุ่ม 5 กลุ่ม สำหรับกรณีที่ 1 และจำนวน 6 กลุ่ม สำหรับกรณีที่ 2, 3 และ 4

ตารางที่ 22 แสดงการเปรียบเทียบค่าความถูกต้องของการทำนายข้อมูล KR ระหว่างตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากวิธีต้นไม้ตัดสินใจของ Hall (HW) และวิธีการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	87.90%	91.84%	**93.37% * +**	93.34% * +
2	10	50%	87.90%	92.09%	93.31% * +	**93.37% * +**
3	10	75%	87.90%	91.56%	93.09% * +	**93.65% * +**
4	1	100%	87.90%	91.50%	**93.53% * +**	93.18% * +

โดยที่ * แสดงว่าผลนั้นมีค่าความถูกต้องของการทำนายมากกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5%

\+ แสดงว่าผลนั้นมีค่าความถูกต้องของการทำนายมากกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งได้มาจากวิธีของ Hall (HW) อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 22 จะทำการคัดเลือกค่าความถูกต้องของการทำนายข้อมูล แต่ละวิธีการจัดกลุ่มข้อมูลที่ให้ค่ามากที่สุดของข้อมูล KR ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 20 และ 21 พบว่าการทำนายโดยวิธี NBCT และ NBCK ให้ค่าที่ใกล้เคียงกันไม่แตกต่างกันนัก และมีค่ามากกว่าทั้งตัวจำแนก NB และตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี Hall (HW) อย่างมีนัยสำคัญ 5%

ตารางที่ 23 แสดงค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล KR สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	50.47 ± 2.77	49.90 ± 2.31	50.18 ± 3.21	**48.59 ± 2.38**	49.04 ± 2.78	49.95 ± 2.53
2	10	50%	50.65 ± 3.34	50.13 ± 2.76	50.26 ± 3.56	**48.72 ± 2.78**	49.01 ± 3.38	49.29 ± 3.26
3	10	75%	50.92 ± 3.10	50.81 ± 2.30	50.33 ± 3.62	**49.22 ± 2.61**	49.63 ± 3.37	49.62 ± 2.75
4	1	100%	50.87 ± 2.81	49.71 ± 2.64	50.66 ± 2.92	49.57 ± 2.94	49.82 ± 3.24	**49.22 ± 2.68**

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 23 พบว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจาก NBCT จะมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 4 กลุ่ม สำหรับกรณีที่ 1, 2 และ 3 และจำนวน 6 กลุ่ม สำหรับกรณีที่ 4

ตารางที่ 24 แสดงค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล KR สำหรับ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการจัดกลุ่มแบบเค-มีนและต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	50.46 ± 2.84	51.27 ± 2.70	50.28 ± 3.48	50.16 ± 3.13	**48.18 ± 3.59**	48.50 ± 3.10
2	10	50%	50.32 ± 2.23	52.28 ± 2.43	49.84 ± 2.87	49.60 ± 2.94	48.18 ± 3.59	**47.98 ± 3.25**
3	10	75%	50.58 ± 2.57	52.89 ± 2.72	49.82 ± 2.82	49.70 ± 3.27	48.50 ± 3.28	**48.26 ± 3.34**
4	1	100%	50.20 ± 2.79	51.91 ± 3.19	49.69 ± 3.05	50.61 ± 3.25	49.33 ± 3.12	**48.91 ± 3.00**

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 24 พบว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธีการ NBCK จะมีค่า RRSE ที่น้อยลงเมื่อจำนวนกลุ่มเพิ่มขึ้นและมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 5 กลุ่ม สำหรับกรณีที่ 1 และจำนวน 6 กลุ่ม สำหรับสำหรับกรณีที่ 2, 3 และ 4

ตารางที่ 25 แสดงการเปรียบเทียบค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ข้อมูล KR ระหว่าง ตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์ที่ใช้น้ำหนัก (WNB) ซึ่งได้มาจากวิธีต้นไม้ตัดสินใจของ Hall (HW) และวิธีการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	60.48 ± 2.93	49.91 ± 3.29	48.59 ± 2.38 * +	**48.18 ± 3.59 * +**
2	10	50%	60.48 ± 2.93	50.32 ± 3.32	48.72 ± 2.78 * +	**48.18 ± 3.59 * +**
3	10	75%	60.48 ± 2.93	50.52 ± 2.75	49.22 ± 2.61 *	**48.26 ± 3.34 * +**
4	1	100%	60.48 ± 2.93	50.54 ± 2.21	49.57 ± 2.94 *	**48.91 ± 3.00 * +**

โดยที่ * แสดงว่าผลนั้นมีค่า RRSE น้อยกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5%

+ แสดงว่าผลนั้นมีค่า RRSE น้อยกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากวิธีต้นไม้ตัดสินใจของ Hall (HW) อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 25 จะทำการคัดเลือกค่า RRSE สำหรับแต่ละวิธีการจัดกลุ่มข้อมูลที่ให้ค่าน้อยที่สุดของข้อมูล KR ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 23 และตารางที่ 24 พบว่า วิธี NBCK ให้ค่า RRSE ที่ดีกว่าวิธี NBCT และมีค่าที่ดีกว่าตัวจำแนก NB และตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี Hall (HW) อย่างมีนัยสำคัญ 5%

สรุปผลการทดลองจากตารางที่ 20 ถึงตารางที่ 25 สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) ของข้อมูล KR จะมีค่าความถูกต้องของการทำนาย และค่า Root Relative Square Error (RRSE) ที่ดีกว่าทั้งตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB) และตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) อย่างมีนัยสำคัญ 5%

เมื่อพิจารณาวิธีการจัดกลุ่มที่เหมาะสมกับข้อมูล KR นั้น วิธีของเค-มีน (NBCK) จะมีค่าความถูกต้องของการทำนาย และค่า RRSE ที่ดีกว่าวิธีการจัดกลุ่มสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative (NBCT)

สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) ของข้อมูล KR จะมีค่าความถูกต้องของการทำนาย และค่า RRSE ที่ดี เมื่อใช้กำหนดจำนวนกลุ่ม 4 กลุ่มสำหรับการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative และ จำนวนกลุ่ม 6 กลุ่ม สำหรับวิธีการจัดกลุ่มแบบ เค-มีน

4. ข้อมูล Wave Form (Wave)

เป็นข้อมูลที่มีขนาดใหญ่ มีจำนวน 5,000 ข้อมูล และมี 40 คุณลักษณะ แต่ละคุณลักษณะเป็นค่าที่ต่อเนื่อง สามารถแบ่งได้เป็น 3 คลาส คือ คลื่นชนิดที่ 1, 2 และ 3 เมื่อนำมาทดสอบกับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) ได้ผลดังตารางที่ 26 – 31 ดังนี้

ตารางที่ 26 แสดงค่าความถูกต้องของการทำนายข้อมูล Wave สำหรับ ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative (TC) และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	82.28%	**82.44%**	82.38%	81.98%	81.30%	80.30%
2	10	50%	81.82%	**82.32%**	82.12%	81.84%	81.28%	80.94%
3	10	75%	82.08%	**82.26%**	81.96%	81.86%	81.64%	80.92%
4	1	100%	81.64%	**82.40%**	81.90%	81.72%	80.16%	79.04%

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 26 พบว่าตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากวิธี NBCT จะให้ค่าความถูกต้องของการทำนายข้อมูลลดลงเมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดเมื่อใช้จำนวนกลุ่ม 2 กลุ่ม

ตารางที่ 27 แสดงค่าความถูกต้องของการทำนายข้อมูล Wave สำหรับ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	81.50%	**82.36%**	82.32%	81.76%	81.80%	81.22%
2	10	50%	82.22%	82.42%	**82.44%**	81.66%	81.68%	81.10%
3	10	75%	81.94%	82.18%	**82.38%**	81.80%	81.64%	81.74%
4	1	100%	81.32%	**82.52%**	82.30%	80.86%	80.42%	80.56%

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

จากตารางที่ 27 พบว่าตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากวิธี NBCK จะให้ค่าความถูกต้องของการทำนายข้อมูลลดลงเมื่อจำนวนกลุ่มมีค่าเพิ่มขึ้น และมีค่าที่ดีที่สุดเมื่อใช้จำนวนกลุ่ม 2 กลุ่ม สำหรับกรณีที่ 1 และ 4 และจำนวน 3 กลุ่ม สำหรับกรณีที่มี 2 และ 3

ตารางที่ 28 แสดงการเปรียบเทียบความถูกต้องของการทำนายข้อมูล Wave ระหว่าง ตัวจำแนกไนอีฟ เบย์ มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) และตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) ซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	80.64%	81.04%	**82.44% * +**	82.36% * +
2	10	50%	80.64%	80.90%	82.32% * +	**82.44% * +**
3	10	75%	80.64%	81.02%	82.26% * +	**82.38% * +**
4	1	100%	80.64%	80.86%	82.40% * +	**82.52% ***

โดยที่ * แสดงว่าผลนั้นมีค่าความถูกต้องของการทำนายมากกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5%

\+ แสดงว่าผลนั้นมีค่าความถูกต้องของการทำนายมากกว่าตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักซึ่งได้มาจากวิธีต้นไม้ตัดสินใจของ Hall (HW) อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 28 จะทำการคัดเลือกค่าความถูกต้องของการทำนาย สำหรับแต่ละวิธีการจัดกลุ่มข้อมูลที่มีค่ามากที่สุดของข้อมูล Wave ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 26 และตารางที่ 27 พบว่าการจัดกลุ่มแบบเค-มีน จะให้ค่าที่ดีกว่าการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ และมีค่ามากกว่าตัวจำแนก NB และตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากวิธี Hall (HW) อย่างมีนัยสำคัญ 5%

ตารางที่ 29 แสดงค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล Wave สำหรับ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักที่ได้มาจากการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ (NBCT)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	65.66 ± 5.83	**62.64 ± 3.85**	63.34 ± 3.95	65.57 ± 3.75	69.62 ± 3.95	74.21 ± 2.36
2	10	50%	65.22 ± 4.99	**63.17 ± 3.85**	64.22 ± 3.84	65.02 ± 4.01	67.53 ± 3.74	70.36 ± 2.70
3	10	75%	64.70 ± 4.92	**63.28 ± 3.85**	64.41 ± 3.80	64.56 ± 4.10	66.10 ± 3.78	68.08 ± 3.05
4	1	100%	64.80 ± 4.87	**63.22 ± 3.96**	64.47 ± 3.71	65.04 ± 4.41	67.00 ± 4.85	69.23 ± 3.82

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 29 พบว่า ตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากวิธี NBCT มีค่า RRSE ที่เพิ่มขึ้นเมื่อจำนวนกลุ่มเพิ่มขึ้นและมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 2 กลุ่ม

ตารางที่ 30 แสดงค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ของข้อมูล Wave สำหรับ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักที่ได้มาจากการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ (NBCK)

วิธีการแบ่งข้อมูล			จำนวนกลุ่ม					
กรณี	ต้นไม้	j	K*	K=2	K=3	K=4	K=5	K=6
1	10	25%	66.52 ± 5.81	62.95 ± 3.57	**62.87 ± 3.93**	66.87 ± 2.85	67.47 ± 2.08	68.71 ± 4.67
2	10	50%	65.19 ± 4.69	**63.33 ± 3.72**	63.42 ± 3.82	65.70 ± 3.76	66.37 ± 2.85	66.61 ± 4.27
3	10	75%	64.42 ± 4.36	**63.35 ± 3.74**	63.48 ± 3.92	64.92 ± 3.86	65.27 ± 3.12	65.86 ± 3.56
4	1	100%	65.37 ± 4.12	**63.40 ± 3.65**	63.83 ± 3.93	66.25 ± 3.88	66.71 ± 2.84	67.26 ± 4.09

โดยที่ j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ
K* คือจำนวนกลุ่มที่ได้จากการพิจารณาค่า BIC (Schwarz's Baysian Criterion)

ในการพิจารณาค่า RRSE ที่ดี คือค่าที่มีค่า RRSE น้อย แสดงให้เห็นว่า ค่าความผิดพลาดของการทำนายมีน้อย จากตารางที่ 30 พบว่า ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากวิธี NBCK มีค่า RRSE เพิ่มขึ้นเมื่อจำนวนกลุ่มเพิ่มขึ้นและมีค่า RRSE ที่ดีที่สุด เมื่อใช้จำนวนกลุ่ม 2 กลุ่ม สำหรับกรณีที่ 2, 3 และ 4 และจำนวน 3 กลุ่ม สำหรับกรณีที่ 1

ตารางที่ 31 แสดงการเปรียบเทียบค่า Root Relative Square Error (RRSE) และส่วนเบี่ยงเบนมาตรฐาน (SD) ข้อมูล Wave ระหว่าง ตัวจำแนกไนอีฟ เบย์ มาตรฐาน (NB), ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) และ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) ซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) (NBCT และ NBCK)

วิธีการแบ่งข้อมูล			NB	Weighted Naive Bayes (WNB)		
กรณี	ต้นไม้	j		Hall (HW)	NBCT	NBCK
1	10	25%	69.38 ± 4.14	65.32 ± 3.89	**62.64 ± 3.85 * +**	62.87 ± 3.93 * +
2	10	50%	69.38 ± 4.14	65.49 ± 4.10	**63.17 ± 3.85 * +**	63.33 ± 3.72 * +
3	10	75%	69.38 ± 4.14	65.53 ± 4.01	**63.28 ± 3.85 * +**	63.35 ± 3.74 * +
4	1	100%	69.38 ± 4.14	66.01 ± 4.03	**63.22 ± 3.96 * +**	63.40 ± 3.65 * +

โดยที่ * แสดงว่าผลนั้นมีค่า RRSE น้อยกว่าตัวจำแนก NB อย่างมีนัยสำคัญ 5%

\+ แสดงว่าผลนั้นมีค่า RRSE น้อยกว่าตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก (WNB) ซึ่งได้มาจากวิธีต้นไม้ตัดสินใจของ Hall (HW) อย่างมีนัยสำคัญ 5%

j คือค่าร้อยละของจำนวนข้อมูลการเรียนรู้ที่จะสุ่มออกมาเพื่อสร้างต้นไม้ตัดสินใจ

NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

จากตารางที่ 31 จะทำการคัดเลือกค่า RRSE สำหรับแต่ละวิธีการจัดกลุ่มข้อมูลที่มีค่ามากที่สุดของข้อมูล Wave ในแต่ละวิธีการแบ่งข้อมูลจากตารางที่ 30 และตารางที่ 31 พบว่าวิธี NBCT มีค่า RRSE ที่ดีกว่าวิธี NBCK และมีค่าที่ดีกว่าตัวจำแนก NB และตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากวิธี Hall (HW) อย่างมีนัยสำคัญ 5%

สรุปผลการทดลองจากตารางที่ 26 ถึงตารางที่ 31 สำหรับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักซึ่งได้มาจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) ของข้อมูล Wave จะมีค่าความถูกต้องของการทำนาย และค่า Root Relative Square Error (RRSE) ที่ดีกว่าทั้งตัวจำแนกไนอีฟ เบย์มาตรฐาน (NB) และตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก (WNB) ซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (HW) อย่างมีนัยสำคัญ 5%

เมื่อพิจารณาวิธีการจัดกลุ่มที่เหมาะสมกับข้อมูล Wave นั้น วิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative จะมีค่าความถูกต้องของการทำนาย และค่า Root Relative Square Error (RRSE) ที่ดีกว่าวิธีของเค-มีน

สำหรับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) ของข้อมูล Wave จะมีค่าความถูกต้องของการทำนาย และค่า RRSE ที่ดี เมื่อกำหนดจำนวนกลุ่ม 2 กลุ่ม ให้กับการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative และการจัดกลุ่มแบบเค-มีน

วิจารณ์

งานวิจัยนี้นำเสนอวิธีการคำนวณน้ำหนักคุณลักษณะโดยการจัดกลุ่มและต้นไม้ตัดสินใจแบบ NBCT และ NBCK สำหรับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักโดยมุ่งเน้นที่จะปรับปรุงประสิทธิภาพให้กับงานวิจัยของ Hall ซึ่งคิดค้นวิธีการคำนวณน้ำหนักคุณลักษณะจากต้นไม้ตัดสินใจสำหรับตัวจำแนกไนอีฟ เบย์ แบบน้ำหนัก และผลการทดลองสามารถเปรียบเทียบให้เห็นดังตารางที่ 32 ซึ่งแสดงผลการดำเนินงานของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ค่าน้ำหนักคำนวณจากการจัดกลุ่มและต้นไม้ตัดสินใจ พบว่าในกรณีของชุดข้อมูล KR และ Wave นั้น อัลกอริทึม NBCT และ NBCK ให้ผลการดำเนินงานที่ดีกว่าอัลกอริทึม NB และ NBH อย่างมีนัยสำคัญ 5% เมื่อพิจารณาจากค่าความถูกต้องของการทำนายและค่า Root Relative Square Error และเมื่อเปรียบเทียบการจัดกลุ่มแบบเค-มีน และ การจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับ วิธีลำดับขั้นแบบ Agglomerative พบว่า ทั้งสองวิธีให้ผลที่ใกล้เคียงกัน

ตารางที่ 32 แสดงการเปรียบผลการทดลองข้อมูลด้วยตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ซึ่งได้จากการจัดกลุ่มและต้นไม้ตัดสินใจ (CDW) กับตัวจำแนกไนอีฟ เบย์ มาตรฐาน (NB) และ ตัวจำแนกไนอีฟ เบย์ แบบน้ำหนักซึ่งได้มาจากต้นไม้ตัดสินใจของ Hall (NBH)

ข้อมูล	จำนวนข้อมูล	จำนวนคุณลักษณะ	จำนวนคลาส	CDW						Naive Bayes (NB)		งานวิจัยของ Hall (NBH)	
				NBCT			NBCK						
				K	Correct	RRSE	K	Correct	RRSE	Correct	RRSE	Correct	RRSE
Bal_sc	625	4	3	2	90.95	58.25	3	90.79	58.84	91.74	65.02*+	91.42	58.08
Crd	1,000	20	2	2	76.50	88.88	2	74.60	89.16	74.60	90.35*+	74.90	89.13
KR	3,196	36	2	4	93.12	48.72	6	93.37	47.98	87.90*+	60.48*+	92.09*+	50.32*+
Wave	5,000	40	3	2	82.38	62.64	3	82.32	62.95	80.64*+	69.38*+	81.04*+	65.32*+

โดยที่ NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

K คือจำนวนกลุ่มที่เหมาะสม

* แสดงว่าผลของการใช้น้ำหนักคุณลักษณะจากวิธี NBCT มีค่าที่ดีกว่าตัวจำแนกที่มีเครื่องหมาย * อย่างมีนัยสำคัญ 5%

\+ แสดงว่าผลของการใช้น้ำหนักคุณลักษณะจากวิธี NBCK มีค่าที่ดีกว่าตัวจำแนกที่มีเครื่องหมาย + อย่างมีนัยสำคัญ 5%

เมื่อพิจารณาจากผลการทดลองของข้อมูล 4 ชุดข้อมูลในตารางที่ 32 พบว่าตัวแปรที่มีผลต่อการเพิ่มประสิทธิภาพและความแม่นยำให้แก่วิธีการคำนวณน้ำหนักคุณลักษณะจากการจัดกลุ่มและต้นไม้ตัดสินใจคือ

1. ขนาดของข้อมูล เมื่อพิจารณาจากผลการทดลองของชุดข้อมูลที่มีขนาดใหญ่ คือชุดข้อมูล KR และข้อมูล Wave พบว่าประสิทธิภาพของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักที่ได้จากการจัดกลุ่มและต้นไม้ตัดสินใจ (NBCT และ NBCK) ซึ่งนำเสนอในการศึกษานี้มีค่ามากกว่าประสิทธิภาพของตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน (NB) และตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากต้นไม้ตัดสินใจของ Hall (NBH) อย่างมีนัยสำคัญ 5% แต่สำหรับชุดข้อมูลที่มีขนาดเล็ก คือผลการทดลองของชุดข้อมูล Bal_sc พบว่า NBCT และ NBCK มี RRSE ที่ด้อยกว่า NBH สำหรับข้อมูลของ Crd พบว่า NBCK มีค่า RRSE ที่ด้อยกว่าอัลกอริทึม NBH

จากที่กล่าวข้างต้น แสดงให้เห็นว่าขนาดของข้อมูลมีผลต่อประสิทธิภาพการทำนายของ NBCT และ NBCK นั่นคือ อัลกอริทึม NBCT และ NBCK ทำงานกับชุดข้อมูลขนาดใหญ่จะให้ประสิทธิภาพที่ดีกว่าชุดข้อมูลที่มีขนาดเล็ก

2. จำนวนคุณลักษณะข้อมูล เมื่อพิจารณาจากชุดข้อมูลที่มีจำนวนคุณลักษณะข้อมูลจำนวนมาก คือชุดข้อมูล KR ประกอบด้วย 36 คุณลักษณะ และข้อมูล Wave ประกอบด้วย 40 คุณลักษณะ พบว่าประสิทธิภาพของตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจ (NBCT และ NBCK) ที่นำเสนอในการศึกษานี้มีค่าเพิ่มขึ้นจากประสิทธิภาพของตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน (NB) และตัวจำแนกไนอีฟ เบย์แบบน้ำหนักจากต้นไม้ตัดสินใจของ Hall (NBH) อย่างมีนัยสำคัญ 5% แต่สำหรับชุดข้อมูลที่มีจำนวนคุณลักษณะน้อย คือผลการทดลองของชุดข้อมูล Bal_sc ซึ่งมี 4 คุณลักษณะ พบว่า NBCT และ NBCK มี RRSE ที่ด้อยกว่า NBH และสำหรับข้อมูลของ Crd ซึ่งมี 20 คุณลักษณะ พบว่า อัลกอริทึม NBCK มีค่า RRSE ที่ด้อยกว่าอัลกอริทึม NBH

จากที่กล่าวข้างต้น แสดงให้เห็นว่าจำนวนคุณลักษณะข้อมูลมีผลต่อประสิทธิภาพการทำนายของอัลกอริทึม NBCT และ NBCK นั่นคือ อัลกอริทึม NBCT และ NBCK ทำงานกับชุดข้อมูลที่มีจำนวนคุณลักษณะข้อมูลมากจะมีประสิทธิภาพที่ดีกว่าชุดข้อมูลที่มีจำนวนคุณลักษณะข้อมูลที่น้อย

3. จำนวนคลาสของข้อมูล เมื่อพิจารณาจากชุดข้อมูล KR ซึ่งข้อมูลแบ่งได้ 2 คลาส จากตารางที่ 23 และ 24 พบว่าประสิทธิภาพและความแม่นยำของไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจ (NBCT และ NBCK) มีค่าที่ดีขึ้นเมื่อจำนวนกลุ่มเพิ่มมากขึ้น เนื่องจากข้อมูลภายในกลุ่มมีการกระจายของข้อมูลในแต่ละคลาสที่ดี นั่นคือ ถึงแม้ว่าจำนวนกลุ่มจะเพิ่มมากขึ้นถึง 6 กลุ่ม แต่ข้อมูลภายในกลุ่มยังคงมีการกระจายข้อมูลครบทั้ง 2 คลาส ทำให้ต้นไม้ตัดสินใจที่สร้างขึ้นให้ค่าน้ำหนักคุณลักษณะที่เหมาะสมกับข้อมูลทดสอบเมื่อนำไปใช้กับตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก แต่เมื่อพิจารณาชุดข้อมูล Wave ซึ่งข้อมูลมีทั้งหมด 3 คลาส จากตารางที่ 29 และ 30 พบว่าประสิทธิภาพและความแม่นยำของไนอีฟ เบย์แบบน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจที่นำเสนอในการศึกษาครั้งนี้มีค่าที่ลดลงเมื่อจำนวนกลุ่มเพิ่มมากขึ้น และมีประสิทธิภาพและความแม่นยำที่ดีเมื่อจัดกลุ่มด้วยจำนวนกลุ่ม 2 กลุ่ม เนื่องจากชุดข้อมูล Wave เมื่อได้รับการแบ่งกลุ่มออกเป็น 2 กลุ่มแล้ว ข้อมูลภายในกลุ่มจะมีการกระจายข้อมูลครบทั้ง 3 คลาส แต่หากได้รับการจัดกลุ่มออกเป็น 3 กลุ่มจะพบว่ามีบางกลุ่มที่ข้อมูลภายในกลุ่มมีการกระจายเพียง 2 คลาส และเมื่อได้รับการจัดกลุ่มออกเป็น 4-6 กลุ่มจะพบว่าบางกลุ่มมีข้อมูลภายในกลุ่มมีการกระจายเพียง 2 คลาสหรือบางกลุ่มพบเพียงคลาสเดียว ซึ่งกรณีนี้จะส่งผลต่อการสร้างต้นไม้ตัดสินใจคือไม่สามารถที่จะสร้างต้นไม้ตัดสินใจได้เนื่องจากข้อมูลนั้นไม่มีความแตกต่างกันของคลาสข้อมูล สำหรับงานวิจัยนี้ ข้อมูลที่อยู่ในกรณีนี้จะกำหนดค่าน้ำหนักคุณลักษณะเท่ากันทุกคุณลักษณะคือ 1 ซึ่งเปรียบเสมือนให้ผลลัพธ์เดียวกับตัวจำแนกไนอีฟ เบย์แบบมาตรฐาน แต่เมื่อเปรียบเทียบกับข้อมูล Crd ซึ่งมี 2 คลาส ในตารางที่ 17 และ 18 พบว่าจำนวนการจัดกลุ่ม 2 กลุ่มมีประสิทธิภาพที่ดีกว่าจำนวน 3 ถึง 6 กลุ่ม ซึ่งอาจเป็นเพราะจำนวนข้อมูล Crd มีค่าน้อย

เมื่อพิจารณาจากผลการทดลองในแต่ละจำนวนกลุ่ม (K) ของชุดข้อมูลที่มีจำนวนคลาสมากพบว่าค่า RRSE มีการเปลี่ยนแปลงมากระหว่างจำนวนกลุ่ม 2 ถึง 6 กลุ่ม คือชุดข้อมูล Bal_sc (3 คลาส) ในตารางที่ 11 และ 12 เช่น จำนวนกลุ่ม 2 กลุ่มมีค่า RRSE 58.25 และจำนวนกลุ่ม 6 กลุ่มมีค่า RRSE 67.43 และชุดข้อมูล Wave (3 คลาส) ดังตารางที่ 29 และ 30 เช่น จำนวนกลุ่ม 2 กลุ่มมีค่า RRSE 62.64 และจำนวนกลุ่ม 6 กลุ่มมีค่า RRSE 74.21 แต่สำหรับชุดข้อมูลที่มีจำนวนคลาสน้อยกว่า พบว่าค่า RRSE มีการเปลี่ยนแปลงเพียงเล็กน้อยคือชุดข้อมูล Crd (2 คลาส) ในตารางที่ 17 และ 18 เช่น จำนวนกลุ่ม 2 กลุ่มมีค่า RRSE 89.16 และจำนวนกลุ่ม 6 กลุ่มมีค่า RRSE 90.27 และชุดข้อมูล KR (2 คลาส) ดังตารางที่ 23 และ 24 เช่น จำนวนกลุ่ม 2 กลุ่มมีค่า RRSE 49.71 และจำนวนกลุ่ม 6 กลุ่มมีค่า RRSE 49.22 ซึ่งเป็นผลมาจากการกระจายของคลาสภายในกลุ่มข้อมูล

จากที่กล่าวข้างต้น แสดงให้เห็นว่าจำนวนคลาสข้อมูลมีผลต่อประสิทธิภาพการทำนายของอัลกอริทึม NBCT และ NBCK นั่นคือ อัลกอริทึม NBCT และ NBCK ทำงานกับชุดข้อมูลที่มีจำนวนคลาสข้อมูลน้อยมีประสิทธิภาพดีกว่าชุดข้อมูลที่มีจำนวนคลาสข้อมูลมากกว่า

เมื่อพิจารณาความซับซ้อนด้านเวลา (Time complexity) ในการเรียนรู้การคำนวณค่าน้ำหนักคุณลักษณะข้อมูลโดยการจัดกลุ่มและต้นไม้ตัดสินใจ สามารถแสดงเปรียบเทียบได้ดังตารางที่ 33

ตารางที่ 33 แสดงการเปรียบเทียบความซับซ้อนด้านเวลาของอัลกอริทึมการคำนวณค่าน้ำหนักคุณลักษณะข้อมูลแบบ CDW และ HW

อัลกอริทึม	ความซับซ้อนด้านเวลา	
	การเรียนรู้ (Training)	การจำแนก (Classification)
ตัวจำแนกไนอีฟ เบย์ มาตรฐาน	O(n*N)	O(n*m)
NBH	$O(n^2*N)$	O(n*m)
NBCK	$O(K*n^2*N)$	O(K*n*m)
NBCT	$O((n*N)^2)$	O(K*n*m)

โดยที่ NBCT คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative และต้นไม้ตัดสินใจ

NBCK คือค่าน้ำหนักคุณลักษณะจากวิธีการจัดกลุ่มแบบเค-มีน และต้นไม้ตัดสินใจ

NBH คือค่าน้ำหนักคุณลักษณะจากวิธีการต้นไม้ตัดสินใจของ Hall

n คือจำนวนคุณลักษณะ

K คือจำนวนกลุ่มของข้อมูลที่ใช้ในการจัดกลุ่ม

N คือจำนวนข้อมูลการเรียนรู้

m คือจำนวนข้อมูลทดสอบ

จากตารางที่ 33 แสดงให้เห็นว่า ตัวจำแนกไนอีฟ เบย์ ที่ใช้น้ำหนักจากวิธีการจัดกลุ่มและต้นไม้ตัดสินใจจะมีค่าซับซ้อนด้านเวลาแตกต่างกันตามการจัดกลุ่ม

สำหรับเวลา (Time Complexity) ในการเรียนรู้ของตัวจำแนกไนอีฟ เบย์แต่ละวิธีมีค่าดังนี้

1. การสร้างตัวจำแนกไนอีฟ เบย์มาตรฐานใช้เวลา O(n*N) (George, 1995)

2. การสร้างตัวจำแนก NBH มี 2 ขั้นตอน คือ 1) การสร้างต้นไม้ตัดสินใจ ใช้เวลาเป็น $O(n^2*N)$ (Su and Zhang,2006) และ 2) การสร้างตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ใช้เวลาเป็น O(n*N) ผลรวมเวลาที่ใช้คือ $O(n^2*N + n*N)$ สรุปเวลาของการสร้างตัวจำแนก NBH มีค่า $O(n^2*N)$

3. การสร้างตัวจำแนก NBCK มี 3 ขั้นตอน คือ 1) การแบ่งกลุ่มข้อมูลด้วยอัลกอริทึมเค-มีน ใช้เวลา O(n*N) (Tan *et al.*, 2006) ต่อมา 2) การสร้างต้นไม้ตัดสินใจให้กับ K กลุ่มข้อมูล ใช้เวลา $O(K*n^2*N)$ และ 3) การสร้างตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ใช้เวลาเป็น O(n*N) ผลรวมเวลาที่ใช้คือ $O((n*N) + (K*n^2*N) + (n*N))$ เท่ากับ $O(n*N*(K*n+2))$ สรุปเวลาของการสร้างตัวจำแนก NBCK มีค่า $O(K*n^2*N)$

4. การสร้างตัวจำแนก NBCT มี 3 ขั้นตอน คือ 1) การแบ่งกลุ่มข้อมูลด้วยการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มและการจัดกลุ่มระดับชั้นแบบ Agglomerative ใช้เวลา $O((n*N)^2)$ (Zhang *et al.*, 1996) ต่อมา 2) การสร้างต้นไม้ตัดสินใจให้กับ K กลุ่มข้อมูล ใช้เวลา $O(K*n^2*N)$ และ 3) การสร้างตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ใช้เวลาเป็น O(n*N) ผลรวมเวลาที่ใช้คือ $O((n*N)^2+(K*n^2*N) + (n*N))$ สรุปเวลาของการสร้างตัวจำแนก NBCK มีค่า $O((n*N)^2)$

การจัดกลุ่มแบบเค-มีน จะใช้เวลา $O(n^2*N)$ ซึ่งมีค่าน้อยกว่าการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerative ใช้เวลาในการสร้างตัวจำแนกเป็น $O((n*N)^2)$ นั่นคือ เมื่อนำมาเปรียบเทียบกับงานวิจัยของ Hall พบว่า ตัวจำแนกไนอีฟ เบย์ ที่ใช้น้ำหนักจากต้นไม้ตัดสินใจของ Hall มีค่าความซับซ้อนของเวลาคือ $O(n^2*N)$ ซึ่งน้อยกว่าตัวจำแนกไนอีฟ เบย์ ที่ใช้น้ำหนักจากการจัดกลุ่มแบบเค-มีนและต้นไม้ตัดสินใจที่ใช้เวลา $O(K*n^2*N)$ ซึ่งแปรตามกับจำนวนกลุ่มที่แบ่งข้อมูล (K) แต่สำหรับการสร้างตัวจำแนกไนอีฟ เบย์ ที่ใช้น้ำหนักจากวิธีการจัดกลุ่มแบบสองขั้นตอนโดยใช้ต้นไม้การจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับชั้นแบบ Agglomerativeและต้นไม้ตัดสินใจที่มีค่าความซับซ้อนของเวลาเท่ากับ $O((n*N)^2)$ ซึ่งเป็นวิธีที่ใช้เวลามากที่สุด

สรุปและข้อเสนอแนะ

สรุป

งานวิจัยนี้นำเสนอวิธีการคำนวณน้ำหนักคุณลักษณะของ ตัวจำแนก Navie Bayes วิธีใหม่ โดยการจัดกลุ่มข้อมูลชุดสอน และสร้างต้นไม้ตัดสินใจจากกลุ่มข้อมูล เพื่อคำนวณนำหนักคุณลักษณะ สำหรับวิธีการจัดกลุ่มที่นำเสนอในงานวิจัยนี้ มี 2 วิธีคือ (1) การจัดกลุ่มแบบสองขั้นตอน ประกอบด้วยการจัดกลุ่มตามคุณลักษณะ (CF Tree) กับวิธีลำดับขั้นแบบ Agglomerative (NBCT) และ (2) วิธีการจัดกลุ่มแบบเค-มีน (NBCK) โดยมีการกำหนดจำนวนกลุ่มที่จะใช้ในงานวิจัย คือ จำนวน 2 กลุ่ม ถึง 6 กลุ่ม และจำนวนกลุ่มที่ได้จากการพิจารณาจากค่า BIC (Schwarz's Baysian Criterion)

ข้อมูลที่ใช้ได้มาจากมหาวิทยาลัยแคลิฟอร์เนีย (University of California, Irvin (UCI)) ซึ่งเป็นชุดข้อมูลมาตรฐาน ซึ่งในงานวิจัยนี้นำมาใช้ 4 ชุดข้อมูล คือ Chess End-Game (KR), Balance Scale Weight& Distance Database (Bal_sc), Wave Form และ German Credit dataset (Crd) สำหรับการประเมินผลการทดลองจะใช้ค่าความถูกต้องของการทำนายและค่า Root Relative Square Error (RRSE) พร้อมทั้งค่าสถิติทดสอบ T-Test ระดับนัยสำคัญ 5% ในการเปรียบเทียบความแตกต่างของ 2 ผลการทดลอง เมื่อนำวิธีการคำนวณคุณลักษณะของตัวจำแนกไนอีฟ เบย์ วิธีต่าง ๆ ได้แก่ NB, NBH, NBCT และ NBCK เปรียบเทียบประสิทธิภาพการทำงาน พบว่าอัลกอริทึม NBCT และ NBCK มีประสิทธิภาพการจำแนกดีกว่าอัลกอริทึม NB และ NBH ในกรณีต่อไปนี้คือ

1. จำนวนข้อมูลที่มีขนาดใหญ่
2. จำนวนคุณลักษณะของข้อมูลที่มีจำนวนมาก
3. จำนวนคลาสของข้อมูลที่มีจำนวนน้อย

เมื่อพิจารณาเวลาการเรียนรู้เพื่อสร้างโมเดลการจำแนกในกรณีของความซับซ้อนของเวลา (Time complexity) พบว่า อัลกอริทึม NBCT และ NBCK ใช้เวลาในการทำงานมากกว่าอัลกอริทึม NBH เพราะงานวิจัยครั้งนี้มีขั้นตอนการทำงานที่เพิ่มขึ้น คือการจัดกลุ่มตามคุณลักษณะ อย่างไรก็ตาม งานวิจัยนี้สามารถใช้เป็นแนวทางในการศึกษาเพื่อพัฒนาวิธีการคำนวณค่าน้ำหนักคุณลักษณะเพื่อใช้ในอัลกอริทึมสำหรับการจำแนกข้อมูลในอนาคตได้

ข้อเสนอแนะ

สำหรับกรณีที่เกิดปัญหาในการสร้างต้นไม้ตัดสินใจ เนื่องจากข้อมูลภายในกลุ่มอยู่คลาสเดียวกัน ทำให้ไม่สามารถสร้างต้นไม้ตัดสินใจได้ ในการศึกษาครั้งนี้จึงกำหนดค่าน้ำหนักทุกคุณลักษณะเท่ากันคือมีค่าเป็น 1 ทำให้โมเดลการจำแนกที่สร้างขึ้นเป็นชุดเดียวกับโมเดลของตัวจำแนกไนอีฟ เบย์มาตรฐาน แนวทางการแก้ปัญหา คือ การใช้วิธีคำนวณน้ำหนักคุณลักษณะโดยวิธีอื่นให้กับกลุ่มข้อมูลที่มีปัญหา เช่น อัลกอริทึมรีลีฟฟ์หรืออัลกอริทึมการเลือกคุณลักษณะบนพื้นฐานความสัมพันธ์ (correlation-based feature selection) เนื่องจากผลการทดลองของน้ำหนักคุณลักษณะที่ได้จากอัลกอริทึมดังกล่าวเมื่อนำมาใช้กับตัวจำแนกไนอีฟ เบย์แบบน้ำหนักพบว่ามีประสิทธิภาพที่ดีกว่าเมื่อเปรียบเทียบกับตัวจำแนกไนอีฟ เบย์มาตรฐาน แต่มีประสิทธิภาพที่น้อยกว่าการคำนวณน้ำหนักจากต้นไม้ตัดสินใจ

ในส่วนของการจัดกลุ่มข้อมูล อาจเพิ่มประสิทธิภาพและความแม่นยำให้กับการทำนายตัวจำแนกไนอีฟ เบย์แบบน้ำหนัก ซึ่งได้จากการคำนวณค่าน้ำหนักจากการจัดกลุ่มและต้นไม้ตัดสินใจได้ โดยอาจจะนำคลาสของข้อมูลไปใช้เป็นคุณลักษณะในการพิจารณาการจัดกลุ่มข้อมูลด้วยซึ่งอาจจะทำให้ข้อมูลภายในกลุ่มมีการกระจายของคลาสข้อมูลได้ดี

เอกสารและสิ่งอ้างอิง

บุญเสริม กิจศิริกุล. 2545. **อัลกอริทึมสำหรับการทำเหมืองข้อมูล.** ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ.

วามิณี นิยากาศ. 2549. **การศึกษาพฤติกรรมและวิเคราะห์การให้บริการด้วยเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอน และ RFM Analysis ในกลุ่มลูกค้าอินเตอร์เน็ตแบงค์กิ้ง.** วิทยานิพนธ์ปริญญาโท, มหาวิทยาลัยเกษตรศาสตร์.

สุรินทร์ นิยมางกูร. 2545. **สถิติเบื้องต้น.** ครั้งที่ 2.กรุงเทพฯ.

Blake, L. and J. Merz. 1998. **UCI repository of machine learning databases.** www.ics.uci.edu/mlearn/MLRepository.html. May 25, 2008.

Breiman, L. 1994. **Bagging Predictor No. 421.**

Cardie, C. and N. Howe. 1997. Improving minority class prediction using case-specific feature weights, pp. 57-65. *In* **Machine Learning 14. ed.** Morgan Kaufmann.

Fraley, C. and A.E. Raftery. 1998. How Many Cluster? Which Clustering Method? Answers Via Model-Based Cluster Analysis, pp. 578-588. *In* **The computer Journal 8. ed.**

Forman, G. and B. Zhang. 2000. Distributed data clustering can be efficient and exact. **ACM SIGKDD Explorations Newsletter** 2 (2): 34 - 38.

George, H.J. and P. Langley. 1995. Estimating Continuous Distributions in Bayesian Classifiers, *In* **Uncertainly in Artificial Intelligence 11. ed.** Morgan Kaufmann Publisher, San Mateo.

Hall, M. 2007. A Decision Tree-Based Attribute Weighting Filter for Naïve Bayes, pp. 120-126. *In* **Knowledge-Based Systems 20. ed.**

Jain, A.K., M.N. Murty and P.J. Flynn. 1999. Data clustering: a review. **ACM Computing Surveys** 31 (3): 164-323.

Kubat, M., D. Flotzinger and G. Pfurtscheller. 1993. Discovering patterns in EEG signals: Comparative study of a few methods, pp. 367-371. *In* **Machine Learning**. Springer-Verleg.

MacQueen, J.B. 1967. Some methods for classification and analysis of multivariate observations, pp. 281-297. *In* **Berkeley Symposium on Mathematical Statistics and Probability 5. ed.** Berkeley, University of California.

Norusis, M.J. 2006. **SPSS 15.0 Guide to Data Analysis.** Prentice Hall.

Mitchell, T.M. 1997. **Machine Learning.** McGraw-Hill, Singapore.

Sinka, M.P. and D.W. Corne. 2002. A large benchmark dataset for web document clustering. **Soft Computing Systems: Design, Management and Applications** 87: 881-890.

Ratanamahatana, C.A. and D. Hunopulos. 2003. Feature selection for the naive Bayesian classifier using decision trees, pp. 475-487. *In* **Applied Artificial Intelligence 17. ed.**

Resampling Stats. **Resampling Stats - XLMiner User Guide.** www.resample.com. April 28, 2008.

Su, J. and H. Zhang. 2006. Full Bayesian Network Classifiers. *In* **Machine Learning 23. ed.**Pittsburgh PA.

Tan, P.N., M. Steinbach and V. Kumar. 2006. **Introduction to Data Mining.** Pearson Education, Inc., USA.

Ward, J.H. **Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association** (58): 236-244.

Wettschereck, D., D.W. Aha and T. Mohri. 1997. A review and empirical comparison of feature weighting methods for a class of lazy learning algorithms, pp. 273-314. *In* **Artificial Intelligence Review 11. ed.**

Witten, I.H. and E. Frank. 2000. **Data mining:Practical Machine Learning Tools and Techniques with Java Implementations.** Morgan Kaufmann, United States of America.

Wiwattanacharoenchai, S. and A. Srivihok. 2003. Understanding online banking in Thailand : cluster analysis of customer usage behavior, *In* **M-business, E-commerce and the impact of broadband on Regional Development and Business Prospects, International Telecommunications Society. Asia-Australasian Regional Conference**.

Zhang, H. and S. Sheng. 2004. Learning Weighted Naive Bayes with Accurate Ranking, pp. 567-570. *In* **IEEE International Conference on Data Mining 4. ed.**

Zhang, T., R. Ramakrishnan and M. Livny. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases, pp. 103-114. *In* **ACM SIGMOD International Conference on Management of Data**.

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นางสาวหทัยชนก กรชี
วัน เดือน ปี ที่เกิด	วันที่ 4 ตุลาคม 2525
สถานที่เกิด	กรุงเทพมหานคร
ประวัติการศึกษา	วท.บ. (ฟิสิกส์) มหาวิทยาลัยเกษตรศาสตร์ บางเขน (พ.ศ. 2547)
ตำแหน่งหน้าที่การงานปัจจุบัน	-
สถานที่ทำงานปัจจุบัน	-
ผลงานดีเด่นและรางวัลทางวิชาการ	1. นำเสนอและตีพิมพ์บทความวิชาการเรื่อง “การปรับน้ำหนักคุณลักษณะด้วยรีลีฟฟ์อัลกอริทึมและต้นไม้ตัดสินใจ สำหรับตัวจำแนกไนอีฟ เบย์ ” ในงานประชุมวิชาการ The 4th International Joint Conference on Computer Science and Software Engineering (JCSSE 2007) จังหวัด ขอนแก่น วันที่ 2-4 พฤษภาคม 2550 2. นำเสนอและตีพิมพ์บทความวิชาการเรื่อง “ReliefF Algorithm and Decision Tree-Based Attribute Weighting for Naive Bayes in Data Classification” ในงานประชุมวิชาการ CoDE 2007 ของสถาบันเทคโนโลยีแห่งเอเชีย วันที่ 10-12 กรกฎาคม 2550
ทุนการศึกษาที่ได้รับ	-