การประยุกต์ใช้ซอฟต์แวร์เวก้าในการจำแนกประเภทข้อมูล

Data Classification Using WEKA Software

นิเวศ จิระวิชิตชัย

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีปทุม วิทยาเขตชลบุรี

E-mail: nivet99@hotmail.com

บทคัดย่อ

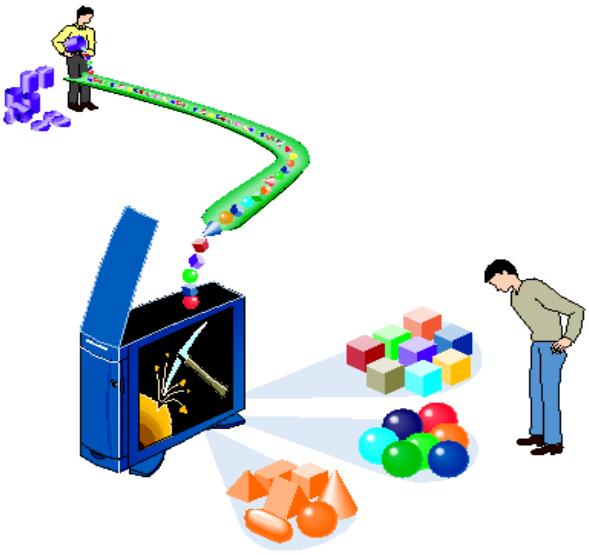
บทความวิชาการนี้ได้นำเสนอวิธีการประยุกต์ใช้ซอฟต์แวร์เวก้าในงานด้านเหมืองข้อมูล โดยมุ่งเน้นไปที่การสร้างแบบจำลองการจำแนกประเภทด้วยต้นไม้การตัดสินใจ เพื่อประโยชน์การในทำนายกลุ่มข้อมูลที่จะเกิดขึ้นในอนาคต ประโยชน์ที่ได้จากบทความนี้คือสามารถนำวิธีต้นไม้การตันสินใจที่นำเสนอไปประยุกต์ใช้ในการจำแนกข้อมูลอื่นๆ เช่น การจำแนกกลุ่มข้อมูลทางการแพทย์ (Medical Classification) ข้อมูลเอกสาร (Text Classification) ข้อมูลเว็บเพจ (Webpage Classification) และระบบตรวจจับการบุกรุก (Intrusion Detection System) เป็นต้น

Abstract

This article presents a methodology of applying data mining through WEKA software. The method focuses on classification and modeling of hidden data in order to make prediction at decision making process. The main benefit of this article is that the decision tree presented in this article can be utilized by other data classifications such a Medical Classification, Text Classification, Webpage Classification and Intrusion Detection System.

1. บทนำ

การทำเหมืองข้อมูล (Data Mining) หรืออาจจะเรียกว่า การค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Databases - KDD) เป็นเทคนิคเพื่อค้นหาภาพแบบ (Pattern) ของจากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ จัดเป็นขบวนการของการดึงเอาความรู้ออกมาจากข้อมูลขนาดใหญ่ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง และ การรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหาภาพแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ [1-2]



ภาพที่ 1 การค้นหาความรู้ในฐานข้อมูล

Data Mining ตามศัพท์ที่ราชบัณฑิตยสถานกำหนดไว้หมายถึง การสกัดหรือวิเคราะห์ ค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมากได้หรือกล่าวอีกนัยหนึ่งคือ ชุดซอฟต์แวร์ (Software) วิเคราะห์ข้อมูลที่ได้ถูกออกแบบมาเพื่อระบบสนับสนุนความต้องการของผู้ใช้ในการค้นหาข้อมูลที่ต้องการจากข้อมูลจำนวนมาก เนื่องจากการทำเหมืองข้อมูลเป็นเทคนิคในการค้นความรู้จากข้อมูลขนาดใหญ่ การทำเหมืองข้อมูลจึงเป็นการรวมเอาศาสตร์ต่างๆ หลายแขนงมารวมไว้ด้วยกันโดยไม่จำกัดวิธีการที่จะใช้ ตัวอย่างศาสตร์ที่ใช้ เช่น เทคโนโลยีฐานข้อมูล (Database Technology) วิทยาศาสตร์สารสนเทศ (Information Science) สถิติ (Statistics) และระบบการเรียนรู้ (Machine Learning) เป็นต้น ซึ่งศาสตร์ต่างๆ เหล่านี้จะทำให้เกิดกระบวนการค้นความรู้ในแบบต่างๆ [1-3]

การทำเหมืองข้อมูลมีขั้นตอนหลักอยู่ 3 ขั้นตอน คือ 1.ขั้นตอนการจัดเตรียมข้อมูล (Preprocessing) ซึ่งการจัดเตรียมข้อมูลนั้นจะต้องทำการคัดข้อมูลที่ไม่เกี่ยวข้องหรือข้อมูลเสีย (Noise Data) ออกจากแหล่งข้อมูลดิบเพื่อให้ได้ข้อมูลที่สัมพันธ์กัน ในขั้นตอนนี้สามารถแบ่งเป็นขั้นตอนย่อยดังนี้ [1-6]

Data cleaning เป็นขั้นตอนในการกำจัดข้อมูลที่เราไม่ต้องการ หรือข้อมูลที่ไม่เป็นประโยชน์ต่อการใช้งานหรือข้อมูลที่มีความผิดพลาด

Data integration เป็นขั้นตอนในการรวบรวมข้อมูลทั้งหมดที่มีจากแหล่งข้อมูลต่างๆ มาไว้ด้วยกัน

Data selection คือการคัดเลือกข้อมูลที่เกี่ยวข้องกับข้อมูลที่จะใช้วิเคราะห์ เลือกข้อมูลที่มีความสัมพันธ์กัน ส่งผลต่อกันและเป็นประโยชน์ต่อการทำนาย

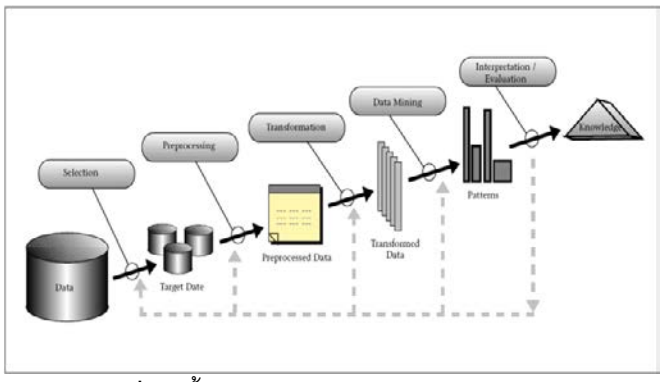
Data transformation เป็นการจัดภาพแบบข้อมูลที่ได้จากขั้นตอนการคัดเลือกข้อมูล ให้มีความเหมาะสมต่อการทำนาย เช่น การจัดระเบียนข้อมูลที่สัมพันธ์กันมาไว้ในระเบียนชุดเดียวกัน หรือการแปลงค่าตัวเลขให้อยู่ในช่วงที่กำหนด (Normalization) เป็นต้น

ขั้นตอนการทำเหมืองข้อมูล (Data Mining) เป็นการนำข้อมูลที่พร้อมแล้วมาสร้างแบบจำลอง โดยขั้นแรกจะต้องทำการเลือกเทคนิคที่เหมาะสมกับภาพแบบชุดข้อมูล พิจารณาปัญหาเช่น ต้องการทำนายประเภทของโรค หรือ ต้องการแบ่งประเภทข้อมูลของโรค หรือต้องการหาปัจจัยที่เกี่ยวข้อง เป็นต้น หลังจากได้เทคนิคที่เหมาะสมแล้วจะทำการสอน (Train) ให้แบบจำลองเรียนรู้ลักษณะของข้อมูลว่า ชุดข้อมูลทั้งหมดมีความสัมพันธ์กันอย่างไร และทิศทางในการวิเคราะห์เป็นอย่างไร โดยในการสอนให้แบบจำลองเรียนรู้นั้น จำเป็นต้องมีการกำหนดพารามิเตอร์ (Parameter) หรือค่าตัวแปรต่างๆให้เหมาะสมซึ่งในการพิจารณาค่าพารามิเตอร์นั้นขึ้นอยู่กับเทคนิคที่เลือกใช้ประสบการณ์ในการวิเคราะห์และการลองผิดลองถูกจากนั้นจึงนำแบบจำลองที่ได้ไปทดสอบหาค่าความผิดพลาดของแบบจำลอง โดยการนำข้อมูลจริงที่เตรียมไว้สำหรับการทดสอบมาป้อนลงในแบบจำลองแล้วดูผลของการทำนายที่ได้

ขั้นตอนสุดท้ายของการทำเหมืองข้อมูล (Post Processing) เป็นขั้นตอนการประเมินผลและนำสิ่งที่ได้มานำเสนอในภาพแบบของการใช้งาน ซึ่งในขั้นตอนนี้หากผลการทดสอบ (Test) ไม่เป็นที่น่าพอใจแล้ว จะต้องทำการจัดภาพแบบข้อมูลใหม่ (หรือเตรียมข้อมูลใหม่) เพื่อให้ได้ ค่าความถูกต้อง (Accuracy)

มากที่สุด หรือค่าความคลาดเคลื่อนในการทำนายน้อยที่สุด (Error) หากค่าความถูกต้องยังน้อยอยู่ หรือความผิดพลาดยังคงมีอยู่ หลังจากเตรียมข้อมูลใหม่อาจจำเป็นต้องเลือกเทคนิคในการทำเหมืองข้อมูลใหม่ ซึ่งการเพิ่มค่าความถูกต้องและลดค่าความคลาดเคลื่อนจะใช้วิธีใดนั้น ขึ้นกับปัญหาที่เกิดขึ้นและในส่วนของขั้นตอนนี้แบ่งเป็นขั้นตอนย่อยๆ ดังนี้ [1-5]

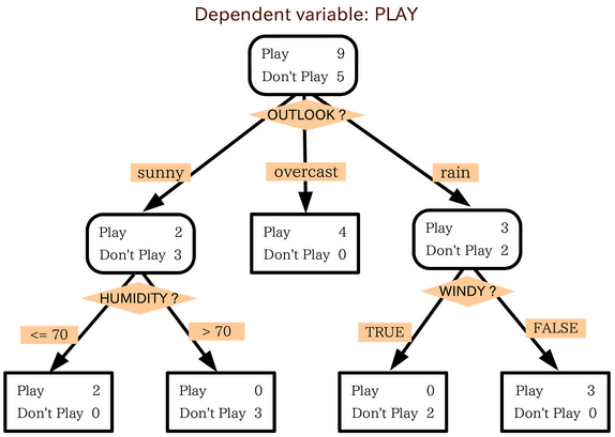
ภาพที่ 2 ขั้นตอนการทำเหมืองข้อมูล



ภาพที่ 3 ขั้นตอนการค้นหาความรู้จากเหมืองข้อมูล

2. การจำแนกประเภท

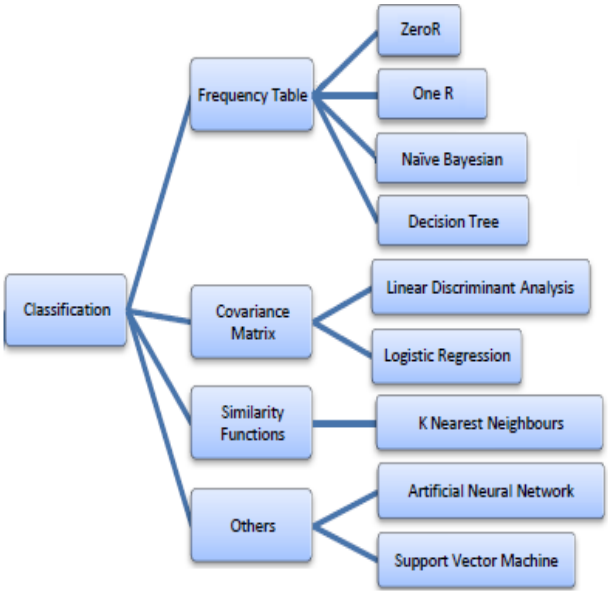
การจำแนกประเภทและการทำนาย (Classification and Prediction) จัดเป็นกระบวนการที่ใช้ในการหาภาพแบบของชุดข้อมูลที่มีความใกล้เคียงกัน หรือเหมือนกันมากที่สุด เพื่อใช้ในการทำนายชุดข้อมูลว่าอยู่ในประเภทใดของชุดข้อมูลที่ได้ทำการแบ่งไว้แล้ว ซึ่งชุดข้อมูลที่แบ่งไว้เกิดจากการเรียนรู้จากชุดข้อมูลที่มีอยู่แล้ว (Training Data) แบบจำลองที่เกิดจากการเรียนรู้ สามารถแสดงได้หลายภาพแบบ เช่น กฎการแบ่ง (Classification Rules, IF-THEN) การคำนวณแบบต้นไม้วิเคราะห์ (Decision Tree) การใช้สูตรทางคณิตศาสตร์ (Mathematical Formula) หรือโครงข่ายประสาทเทียม เป็นต้น ในส่วนของการทำต้นไม้วิเคราะห์ จะแสดงออกมาในลักษณะของแผนภูมิโครงสร้างต้นไม้ ซึ่งก้านของต้นไม้จะแสดงถึงความรู้ที่ได้ และใบไม้จะแสดงถึงประเภทชุดข้อมูลที่ถูกแบ่งออกมา แผนภูมิต้นไม้สามารถแปลงเป็นกฎการแบ่งได้ง่ายเพราะลักษณะของแผนภูมิสามารถเข้าใจได้ง่าย [6-8]



ภาพที่ 4 แผนภูมิต้นไม้

ในส่วนของโครงข่ายใยประสาทเทียมนั้น จะแสดงในลักษณะของการเชื่อมต่อระหว่างหน่วยที่เกิดขึ้น การทำการแบ่งประเภทนั้นมักจะใช้ประโยชน์ร่วมกับการทำนายโดยเฉพาะข้อมูลที่เป็นตัวเลข เราจึงอาจมองได้ว่าการทำนายเป็นการบอกถึงค่าตัวเลขและการบ่งบอกประเภทของข้อมูลนั้นในลักษณะของการดูแนวโน้ม (Trends) ที่จะเกิดขึ้น ตัวอย่างเทคนิคของการแบ่งประเภทและการทำนายได้แก่ ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) เนอีฟเบย์ (Naïve-Bayes) ต้นไม้การตัดสินใจ (Decision Tree) การคำนวณแบบพันธุกรรม (Genertic

Algorithm) และโครงข่ายประสาทเทียม (Neural Network) เป็นต้น [6-8]



ภาพที่ 5 วิธีการจำแนกประเภท

3. ซอฟต์แวร์เวก้า

โปรแกรมเวก้า (Waikato Environment for Knowledge Analysis: Weka) [9-11] เริ่มพัฒนามาตั้งแต่ปี 1997 โดยมหาวิทยาลัย Waikato ประเทศนิวซีแลนด์ เป็นซอฟต์แวร์สำเร็จรูปประเภทฟรีแวร์ อยู่ภายใต้การควบคุมของ GPL License ซึ่งโปรแกรมเวก้าได้ถูกพัฒนามาจากภาษาจาวาทั้งหมด ซึ่งเขียนมาโดยเน้นกับงานทางด้านการเรียนรู้ด้วยเครื่อง (Machine Learning) และ การทำเหมืองข้อมูล (Data Mining)โปรแกรมจะประกอบไปด้วยโมดูลย่อย ๆ สำหรับใช้ในการจัดการข้อมูล และ เป็นโปรแกรมที่สามารถใช้งาน Graphic User Interface (GUI) และใช้คำสั่งในการให้ซอฟต์แวร์ประมวลผล และ สามารถรันได้หลายระบบปฏิบัติการ และสามารถพัฒนาต่อยอดโปรแกรมได้ เป็นเครื่องมือที่ใช้ทำงานในด้านการทำเหมืองข้อมูลที่รวบรวมแนวคิดอัลกอริทึมมากมาย ซึ่งอัลกอริทึมเหล่านั้นสามารถเลือกใช้งานโดยตรงได้จาก 2 ทางคือจากชุดเครื่องมือที่มีอัลกอริทึมมาให้หรือเลือกใช้จากอัลกอริทึมที่ได้ เขียนเป็นโปรแกรมลงไปเป็นชุดเครื่องมือเพิ่มเติม และชุดเครื่องมือมีฟังก์ชันสำหรับการทำงานร่วมกับข้อมูล ความสามารถของซอฟต์แวร์เวก้าสนับสนุนเกี่ยวกับการทำเหมืองข้อมูล (Data Mining) [9-11]

การเตรียมข้อมูล (Data Preprocessing) การทำเหมืองข้อมูลด้วย เทคนิคการจำแนกข้อมูล (Classification) การทำเหมืองข้อมูลด้วยเทคนิคการจัดกลุ่ม (Clustering) การทำเหมืองข้อมูลด้วยเทคนิคการวิเคราะห์ความสัมพันธ์ (Associating) เทคนิคการคัดเลือกข้อมูล (Selecting Attributes) เทคนิคการนำเสนอข้อมูลด้วยรูปภาพ (Visualization) โดยโลโก้ของซอฟต์แวร์เวก้าเป็นรูปนกท้องถิ่นของประเทศนิวซีแลนด์ [9-11] ดังภาพ

ภาพที่ 6 โปรแกรมเวก้า

การทำเหมืองข้อมูลการจัดจำแนกประเภทจัดเป็นการสร้างตัวแบบ Classifier ที่สามารถแบ่งแยกข้อมูล ออกตามคลาสหรือลักษณะประจำเป้าหมายที่กำหนด โดยตัวแบบจำลองที่ต้องการอาจเลือกจากวิธีการดังต่อไปนี้

- Bayes ใช้หลักของเบย์หรือตัวแบบเชิงความน่าจะเป็น

- Functions ตัวแบบในรูปของฟังก์ชัน

- Lazy ตัวแบบที่เก็บตัวอย่างการตัดสินใจเกิดเมื่อตัวอย่างใหญ่ถูกนำเข้าเท่านั้น

- Meta การทำตัวแบบให้ดีขึ้นโดยการเรียนข้อมูลเมต้า

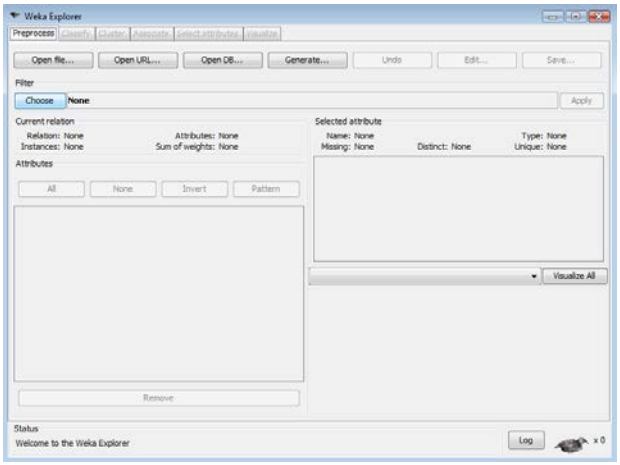
- Misc วิธีการสร้างตัวแบบวิธีอื่น

- Trees การสร้างตัวแบบโดยใช้ต้นไม้

- Rules การสร้างตัวแบบโดยใช้กฎ

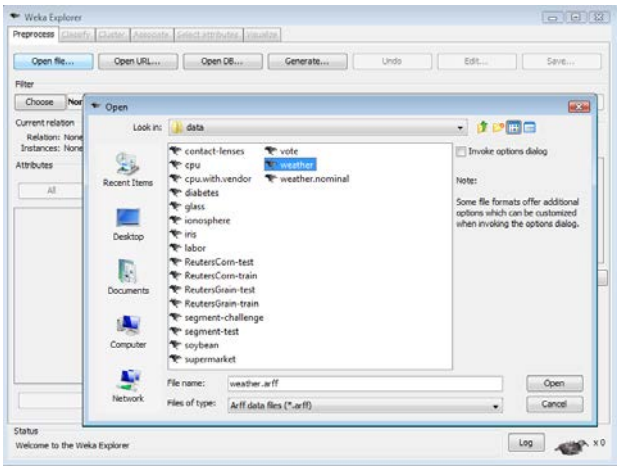
4. การใช้เวก้าในงานจำแนกประเภท

ตัวอย่างของบทความวิชาการนี้ เราจะประยุกต์ใช้โปรแกรมเวก้า โดยใช้ขั้นตอนวิธีการสร้างตัวแบบโดยใช้ ต้นไม้การตัดสินใจ (Decision Tree) [12] ในการจำแนกประเภทการเล่นกีฬาตามสภาพอากาศ ว่าควรเล่นหรือไม่เล่น ดังตัวอย่าง เริ่มการทำงานของซอฟต์แวร์ Weka ที่ C:\Program Files\Weka-3.6 หรือ ไปที่ Start ▶ All programs ▶ Weka 3.6.3



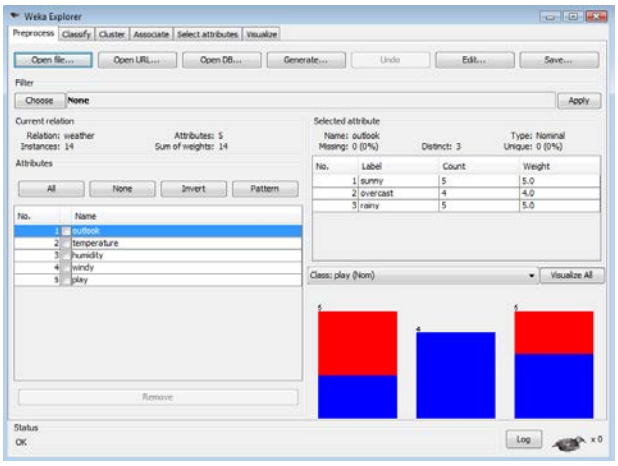
ภาพที่ 7 เปิดแฟ้มข้อมูล

เลือก Weka เปิดโมดูล Explorer คลิกที่ Open file เปิดแฟ้ม weather.arff ที่C:\Program Files\Weka-3.6\data

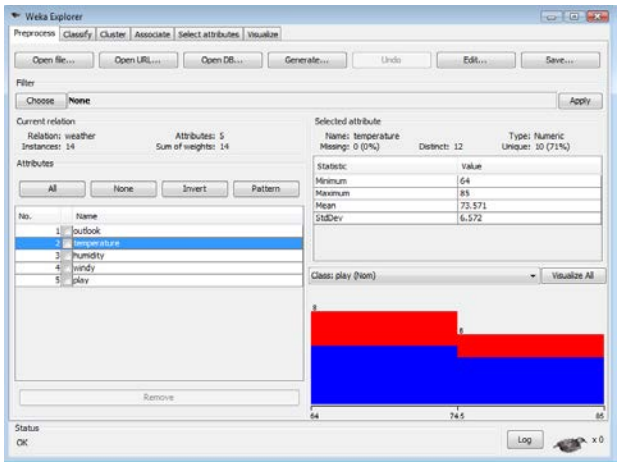


ภาพที่ 8 เลือกแฟ้มข้อมูล weather

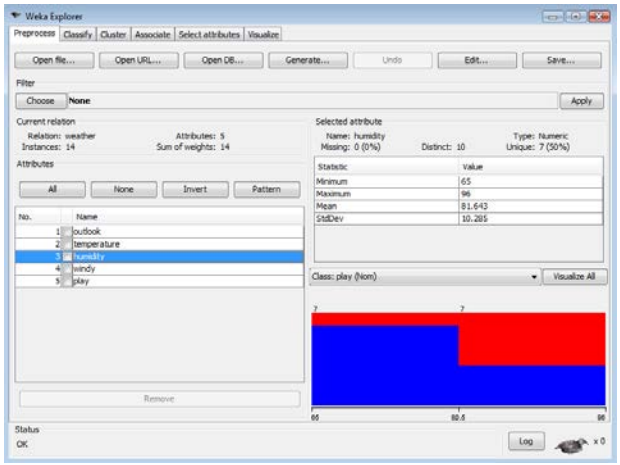
ทำการตรวจสอบค่าความแตกต่างกันในแต่ละคุณลักษณะ (Distinct) ในฐานข้อมูลสภาพอากาศ weather ค่าความแตกต่างกันของคุณลักษณะ outlook มี 3 ค่า temparature มี 12 ค่า humidity มี 10 ค่า windy มี 2 ค่า play มี 2 ค่า



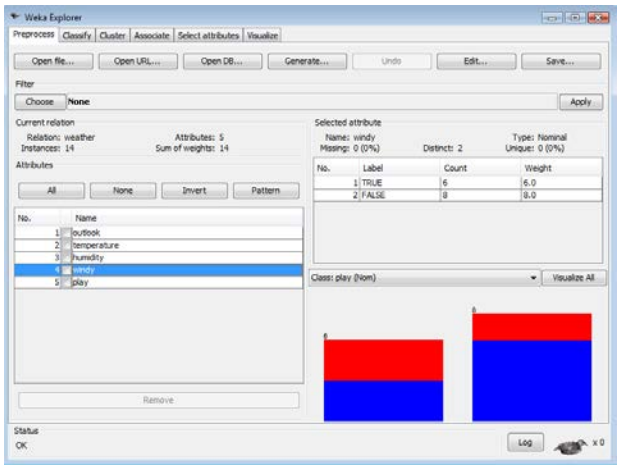
ภาพที่ 9 ค่าความแตกต่างกันของคุณลักษณะ outlook



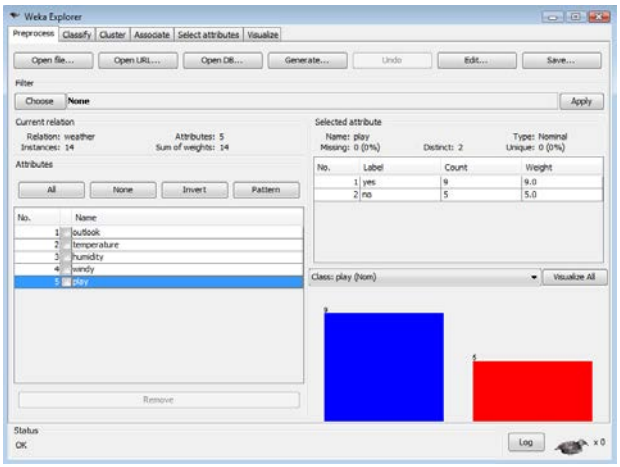
ภาพที่ 10 ค่าความแตกต่างกันของ temparature



ภาพที่ 11 ค่าความแตกต่างกันของคุณลักษณะ humidity



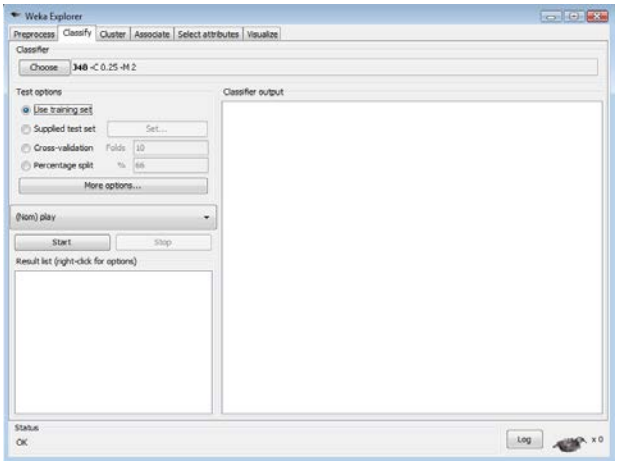
ภาพที่ 12 ค่าความแตกต่างกันของคุณลักษณะ windy



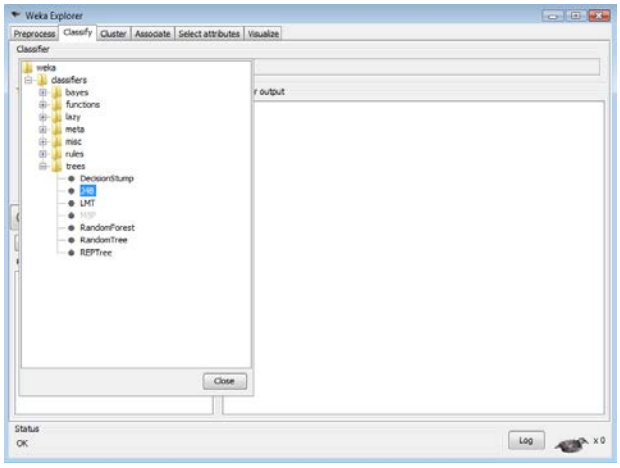
ภาพที่ 13 ค่าความแตกต่างกันของคลาส play

5. การใช้เวก้าในงานจำแนกประเภท

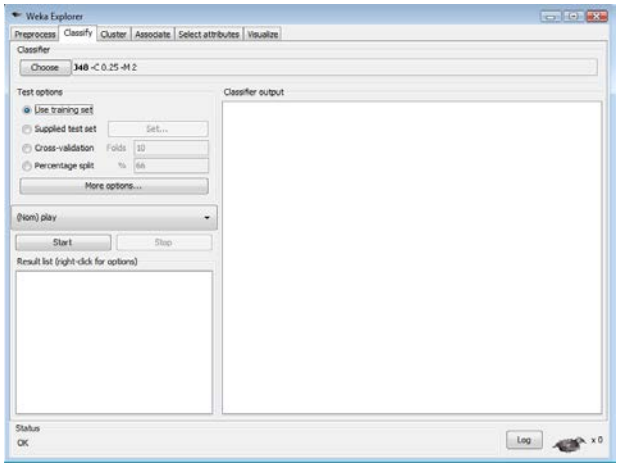
ในบทความวิชาการนี้ จะนำเสนอการทดสอบการจำแนกประเภทจากกลุ่มข้อมูล weather ที่นำมาเรียนรู้เพื่อสร้างโมเดลการตัดสินใจ โดยเลือกแถบ Classify เลือก Choose แล้วเลือกต้นไม้การตันสินใจ tree ▶ j48 (Decision Tree)



ภาพที่ 14 เลือกแถบ Classify เลือกปุ่ม Choose

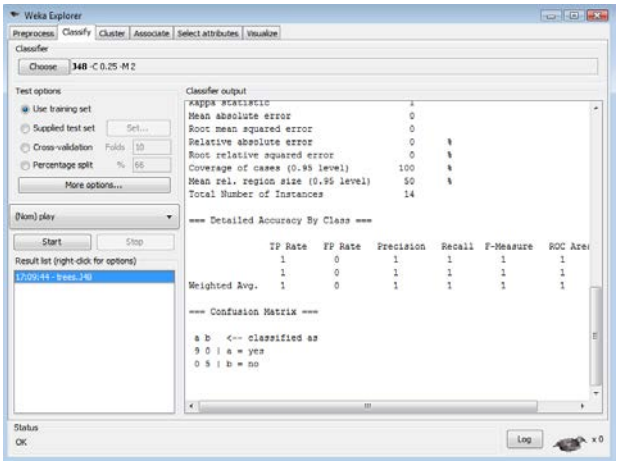


ภาพที่ 15 เลือกต้นไม้การตันสินใจ j48



ภาพที่ 16 เลือก test option

ให้กำหนดตัวเลือก use training set เพื่อใช้ทุกตัวอย่างในการสร้างต้นไม้การตัดสินใจ กำหนดคลาสเลือกคุณลักษณะเป้าหมายที่ต้องการ โดยปรกติคุณลักษณะสุดท้าย (Class) จะถูกเลือกโดยอัตโนมัติ กดปุ่ม Start เพื่อเริ่มสร้างต้นไม้ จะได้ผลลัพธ์ตามภาพ



ภาพที่ 17 แสดง Confusion matrix

รายงานผลลัพธ์ของตัวแบบกับกลุ่มข้อมูลเรียนรู้ (training set) เมื่อพิจารณาถึง Classification Confusion matrix แสดงค่าที่ได้จากตัวแบบ กับค่าจริง โดยผลลัพธ์ที่ดีต้องไม่มีค่านอก diagonal

6. ผลการวิเคราะห์ด้วยซอฟต์แวร์เวก้า

เมื่อทำการสร้างแบบจำลองด้วยเทคนิควิธีต้นไม้การตันสินใจ ในการเรียนรู้กลุ่มข้อมูลสภาพอากาศ weather ส่วนของ Run Information สามารถสรุปได้ว่า [13-15]

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     weather
Instances:    14
Attributes:   5
              outlook
              temperature
              humidity
              windy
              play
Test mode:    evaluate on training data
```

ภาพที่ 18 แสดงส่วน Run Information

- Scheme : เทคนิคที่ใช้คือการจำแนกประเภท

- Relation : ข้อมูลที่ใช้นำเข้ามีชื่อว่าอะไร Instances : จำนวนแถวในข้อมูล

- Attributes : จำนวนคอลัมน์ในข้อมูล และชื่อของแต่ละคุณลักษณะ

-Test mode : การทดสอบประสิทธิภาพของโมเดลที่ได้จากการจำแนกประเภท

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
outlook = overcast: yes (4.0)
outlook = rainy
|   windy = TRUE: no (2.0)
|   windy = FALSE: yes (3.0)

Number of Leaves  :     5

Size of the tree :      8


Time taken to build model: 0.27 seconds
```

ภาพที่ 19 แสดงส่วน Classifier Model

ส่วนของ Classifier model คือส่วนของโมเดลที่สร้างได้ ซึ่งจะแตกต่างกันออกไปตามเทคนิควิธีที่เลือกใช้งานจำแนกประเภทข้อมูล ในบทความนี้ใช้เทคนิคการทำ classification ด้วยวิธี J48 ซึ่งเป็น Decision tree แบบหนึ่ง ผลที่แสดงในส่วนนี้จึงเป็นลักษณะของ tree แต่เขียนให้อยู่ในรูปของ text ซึ่งจาก decision tree นี้เราอาจจะสามารถการแปลผลโดยแปลงให้เป็นกฎที่เป็นโมเดลได้ เช่น ถ้าค่า outlook = sunny และ humidity <= 75 ก็จะตอบว่า play = yes หรือถ้า outlook = overcast จะตอบว่า play = yes หรือ outlook = rainy และ windy = FALSE จะตอบว่า play = yes เป็นต้น ส่วนของ Evaluation คือส่วนที่วัดประสิทธิภาพของโมเดลในการ classify ที่สร้างขึ้นมาได้ ซึ่งมีค่าที่สำคัญๆ อยู่ 2 ค่า คือ Correctly Classified Instances ส่วนนี้บอกว่าจากข้อมูลที่มี 14 instance นั้น มีการทำนายข้อมูลถูกต้อง 14 instance (row) หรือคิดเป็น 100% ของข้อมูลทั้งหมด

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances          14               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances               14
```

ภาพที่ 20 แสดงส่วน Evaluation on training set

Incorrectly Classified Instances ส่วนนี้บอกว่าจากข้อมูลที่มี 14 instance นั้น มีการทำนายข้อมูลไม่ถูกต้อง 0 instance (row) หรือคิดเป็น 0% ของข้อมูลทั้งหมด

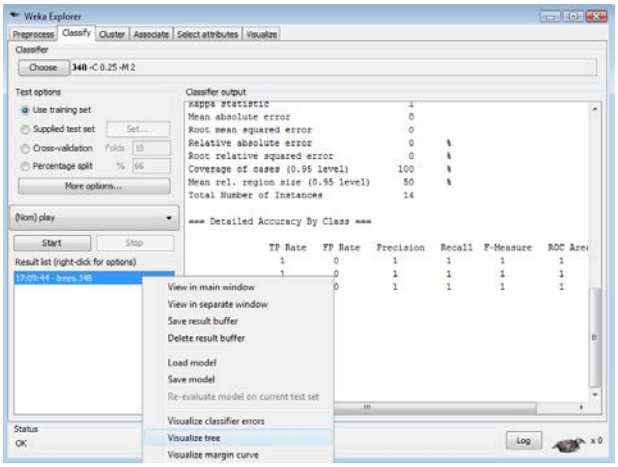
```
=== Confusion Matrix ===

 a b   <-- classified as
 9 0 | a = yes
 0 5 | b = no
```

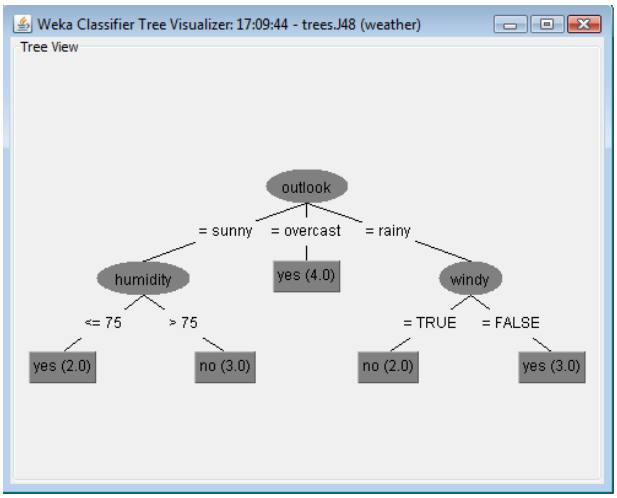
ภาพที่ 21 แสดงส่วน Confusion matrix

ส่วนของ Confusion Matrix คือ ส่วนที่ทำให้เราเห็นรายละเอียดของผลการทำนายของข้อมูลแต่ละคลาสได้ดีขึ้น ค่าในคอลัมน์ คือ ค่าที่ได้จากการทำนายด้วยเทคนิค J48 ส่วนค่าในแถวจะเป็นส่วนของค่าที่

เป็นคำตอบหรือเฉลยของคลาสนั้น จากตารางนี้เราจะเห็นว่าข้อมูลที่จริงๆแล้วมีค่า play=yes และโมเดลของเราทำนายถูกว่ามีค่า play=yes นั้นมีจำนวนทั้งหมด 9 instance ข้อมูลเหล่านี้ คือ ข้อมูลที่โมเดลทำนายถูกเช่น ข้อมูลใน instance ที่มีค่า play=yes แต่ถ้าจริงๆ แล้วข้อมูลมีค่า play=yes แต่โมเดลการทำนายดันไปตอบว่าค่า play=no นั้นมีจำนวน 5 instance นั้นคือข้อมูลที่โมเดลทำนายผิดสำหรับคลาสที่ตอบว่า play=yes นั่นเอง โดยเราสามารถดูแผนแสดงแผนผังต้นไม้ได้โดย คลิกขวาที่ชื่อ Model และเลือก Visualize Tree



ภาพที่ 22 แสดงภาพแผนผังต้นไม้



ภาพที่ 23 แผนผังต้นไม้ของข้อมูล weather

6. บทสรุป

บทความวิชาการนี้ได้นำเสนอวิธีการประยุกต์ใช้ซอฟต์แวร์เวก้าในงานด้านเหมืองข้อมูล โดยมุ่งเน้นไปที่การสร้างแบบจำลองการจำแนกประเภทข้อมูล โดยใช้ต้นไม้การตัดสินใจมาเป็นเครื่องมือ เพื่อประโยชน์การในทำนายกลุ่มข้อมูลที่จะเกิดขึ้นในอนาคต อย่างไรก็ตามการจำแนกประเภทในเหมืองข้อมูลนั้น มีหลากหลายวิธี ซึ่งแต่ละวิธีก็มีจุดเด่นแตกต่างกันไป แต่ที่บทความนี้นำเสนอวิธีต้นไม้การตัดสินใจ เพราะง่ายต่อการศึกษาและทำความเข้าใจ จำแนกข้อมูลได้รวดเร็วและมีประสิทธิภาพ ตลอดจนสามารถนำโมเดลไปพัฒนาเชิงโปรแกรมประยุกต์ได้ง่าย

นอกจากนั้นประโยชน์ที่ได้จากบทความนี้คือสามารถนำวิธีต้นไม้การตันสินใจที่นำเสนอไปประยุกต์ใช้ในการจำแนกข้อมูลอื่นๆ เช่น การจำแนกกลุ่มข้อมูลทางการแพทย์ (Medical Classification) ข้อมูลเอกสาร (Text Classification) ข้อมูลเว็บเพจ (Webpage Classification) ระบบตรวจจับการบุกรุก (Intrusion Detection System) เป็นต้น

7. เอกสารอ้างอิง

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, "Practical Machine Learning Tools and Techniques" , 3rd Edition, Morgan Kaufman Publishers, 2011.

[2] Han and Kamber, “Data Mining Concepts and Techniques” , San Francisco, Morgan Kaufmann Publishers, 2006.

[3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction to Data Mining", Addison Wesley Publishers, 2005.

[4] Fayyad, U., Grinstein, G. and Wierse, A., "Information Visualization in Data Mining and Knowledge Discovery", Morgan Kaufmann Publishers, 2001.

[5] Fayyad, U., Piatetsky-shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases", AI Magazine, Volume 17, 1996.

[6] อุกฤษ ปัจฉิม. "การประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการทำนายระดับน้ำสูงสุด", วิทยานิพนธ์ วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมทรัพยากรน้ำ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี, 2546.

[7] นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ,“การศึกษาทดลองเทคนิคการลดคุณลักษณะและอัลกอริทึมการจัดหมวดหมู่ของเอกสารภาษาไทย”, วารสารวิทยาศาสตร์ลาดกระบัง, ปี 2552.

[8] นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ, “การเปรียบเทียบการคำนวณน้ำหนักดัชนีสำหรับบอัลกอริทึมการจัดหมวดหมู่เอกสารภาษาไทย”, วารสารวิทยาศาสตร์ลาดกระบัง, ปี 2553.

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

[10]David Scuse. Peter Reutemann, "WEKA Experimenter Tutorial for Version 3-5-8", WEKA Manual, 2008.

[11]Richard Kirkby, Eibe Frank, Peter Reutemann, "WEKA Explorer User Guide for Version 3-5-8", WEKA Manual, 2008.

[12] Quinlan, J. R., "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[13] Bamshad Mobasher, "Classification via Decision Trees in WEKA",http://maya.cs.depaul.edu/classes/weka/classify.html

[14] Croce Danilo, Roberto Basili, "Decision tree algorithm short Weka tutorial", http://art.uniroma2.it/basili/MLWM09/002_DecTree_Weka.pdf

[15] Zdravko Markov, "An Introduction to the WEKA Data Mining System", www.cs.ccsu.edu/~markov/weka-tutorial.pdf