CSC662 ปฏิบัติการ ๑๑

การใช้ผังการไหลของความรู้ Knowledge Flow

เขียนโดย ผศ. ดร. กรุง สินอภิรมย์สราญ

ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

เนื้อหาที่ครอบคลุม

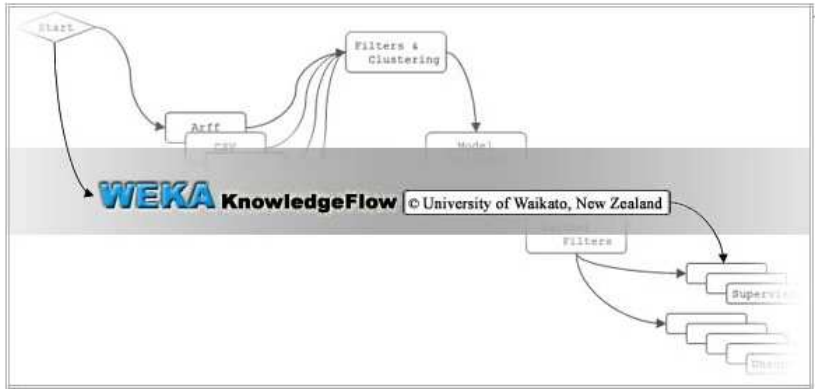
- แนะนำซอฟต์แวร์ Weka
- ฐานข้อมูลเชิงสัมพันธ์และการเตรียมข้อมูล
- สถิติและการทำเหมืองข้อมูลแบบอธิบาย
- การทำเหมืองข้อมูลแบบกฎเชื่อมโยง
- การทำเหมืองข้อมูลแบบจำแนกประเภท
- การทำเหมืองข้อมูลแบบการวิเคราะห์การเกาะกลุ่ม
- การใช้ผังการไหลของความรู้ Knowledge flow

นิยามผังการไหลของความรู้

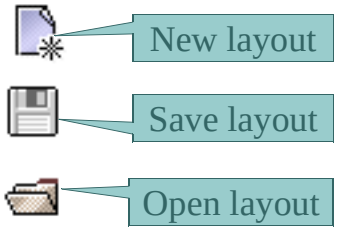
- ผังการไหลของความรู้ คือแผนภาพที่แสดงถึงการได้มาของความรู้ (Knowledge) โดยผ่านกระบวนการ/ขั้นตอนวิธี/การแสดงผลที่ใช้ในการทำเหมืองข้อมูล
- แผนภาพที่สร้างใช้สัญลักษณ์ (Icon) ที่สื่อถึงกระบวนการ/ขั้นตอนวิธี/การแสดงผลหนึ่งลักษณะหรือหนึ่งแบบ
- เส้นที่เชื่อมโยงระหว่างสัญลักษณ์แสดงการไหลของข้อมูล (data) ที่ผ่านกระบวนการ (icon) จนถึงความรู้ที่ได้

ตัวอย่าง การไหลของข้อมูลเพื่อให้ได้ความรู้ DataSources → Filter → Classifier → Evaluator → Visualization

หน้าจอเริ่มต้นของผังการไหลของความรู้

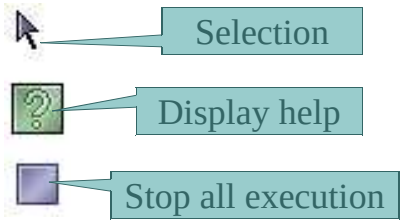


รายการของผังการไหลของความรู้



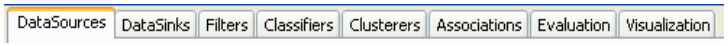
- New layout สร้างผังการไหลใหม่
- Save layout เก็บผังที่สร้างไว้ใน Knowledge Flow Layout บันทึกลง แฟ้มข้อมูลเพื่อนำกลับมาใช้ใหม่
- Open layout เปิดแฟ้มข้อมูลที่เก็บผังที่สร้างแล้ว เพื่อนำกลับมาใช้ใหม่

รายการของผังการไหลของความรู้



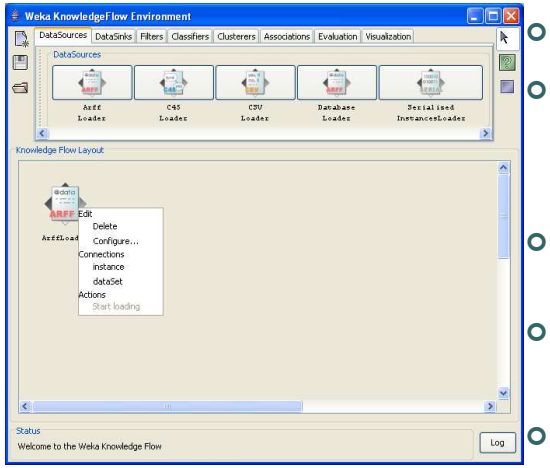
- Selection เปลี่ยนเมาส์ให้เป็นตัวชี้เพื่อเลือกภาพสัญลักษณ์หรือเส้นเชื่อม
- Display help แสดงข้อความอธิบายการใช้เครื่องมือของผังการไหลของความรู้
- Stop all execution หยุดการประมวลผลทุกอย่างที่เกิดขึ้น

ส่วนประกอบหลักของผังการไหลของความรู้



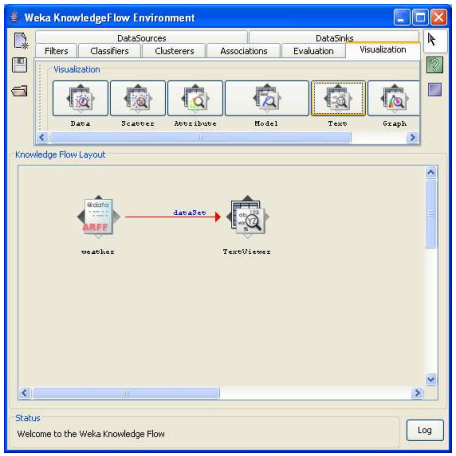
- DataSources: กำหนดแหล่งข้อมูลที่อ่านเข้าผัง
- DataSinks: กำหนดการบันทึกข้อมูลหรือจุดสุดท้ายของกระบวนการ
- Filters: ขั้นตอนการจัดการเตรียมข้อมูล
- Classifiers: การสร้างตัวแบบและวิธีการในการจัดจำแนกประเภท
- Clusterers: การใช้ขั้นตอนวิธีการวิเคราะห์การเกาะกลุ่ม
- Associations: การใช้ขั้นตอนวิธีการหากฎเชื่อมโยง
- Evaluation: ประเมินและแบ่งเซตข้อมูลออกเป็นส่วน ๆ
- Visualization: สำหรับแสดงผลลัพธ์ด้วยภาพนามธรรม

ตัวอย่างการสร้างผังอ่านข้อมูล



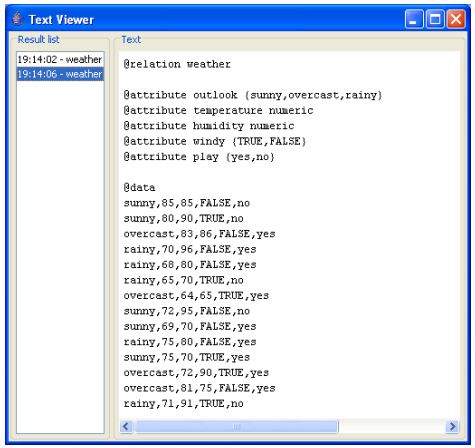
- เริ่มจากเลือกแถบ DataSources
- เลือกสัญลักษณ์ ArffLoader เมาส์เปลี่ยนเป็นเครื่องหมายกากบาท
- กดเมาส์ในบริเวณ Knowledge Flow Layout
- กดเมาส์ปุ่มขวาที่ ArffLoader เลือก Configure…
- เลือกแฟ้มที่ชื่อ weather.arff

ตัวอย่างผังการไหลที่แสดงข้อความของข้อมูล



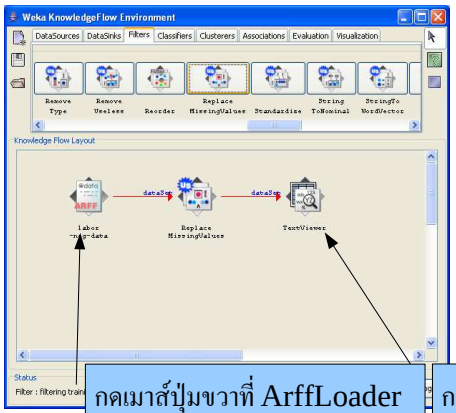
- เลือกแถบ Visualization
- กดเมาส์ที่ Text Viewer เมาส์ เปลี่ยนเป็นเครื่องหมายกากบาท
- กดเมาส์ปุ่มขวาเพื่อเชื่อม ArffLoader ไปยัง TextViewer โดยเลือก dataSet บนเมนูของ weather.arff
- กดเมาส์ปุ่มขวาที่ ArffLoader โดยเลือก Start loading ใต้รายการ Action

การแสดงผลของที่สัญลักษณ์ภาพนามธรรม



- หลังจาก เลือก Start loading ใต้รายการ Action ของ ArffLoader
- ชื่อแฟ้มที่อ่านจะปรากฏใต้ภาพ ArffLoader
- แสดงข้อความโดยเลือก Show results ภายใต้รายการใน TextViewer โดยกดเมาส์ปุ่มขวา
- ผลลัพธ์ที่ได้แสดงทางภาพด้านซ้าย

การเพิ่มขั้นตอนในการกรอง Missing value

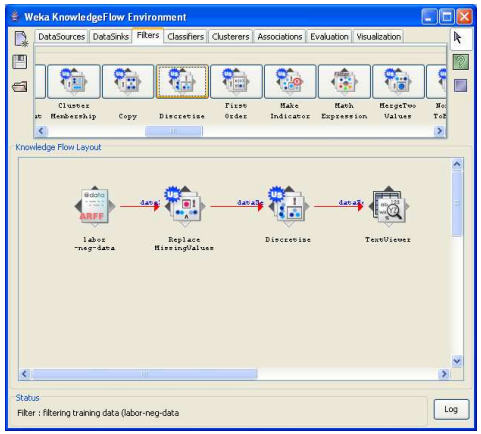


- เริ่มจาก DataSources โดยใช้ ArffLoader
- เลือก Configure… แล้วเลือกแฟ้ม labor.arff
- เลือกแถบ Filters ที่เรียก Replace MissingValues เพื่อเติมค่าที่หายไป
- เลือกแถบ Visualization แล้วเลือก TextViewer เพื่อแสดงผลลัพธ์

กดเมาส์ปุ่มขวาที่ ArffLoader แล้วเลือก Start loading

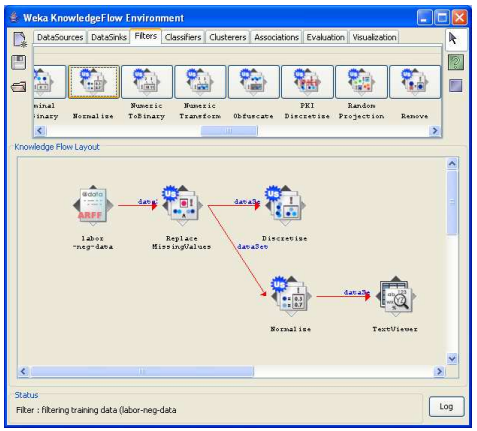
กดเมาส์ปุ่มขวาที่ TextViewer แล้วเลือก Show results

การเพิ่มขั้นตอน Discretization



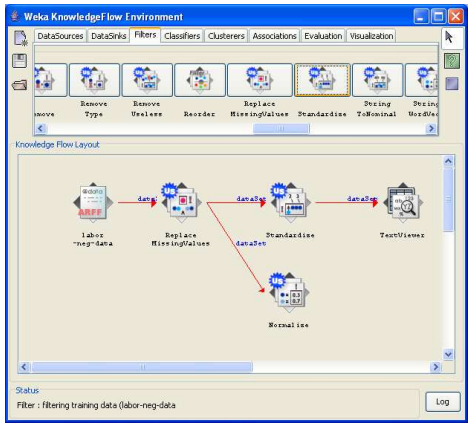
- เริ่มจาก DataSources โดยใช้ ArffLoader
- เลือก Configure… แล้วเลือกแฟ้ม labor.arff
- เลือกแถบ Filters แล้วเลือก Discretize เพื่อเปลี่ยนตัวแปรที่มีค่าต่อเนื่องเป็นตัวแปรที่มีค่าไม่ต่อเนื่อง
- เลือก TextViewer จากแถบ Visualization

การเพิ่มขั้นตอน Normalization



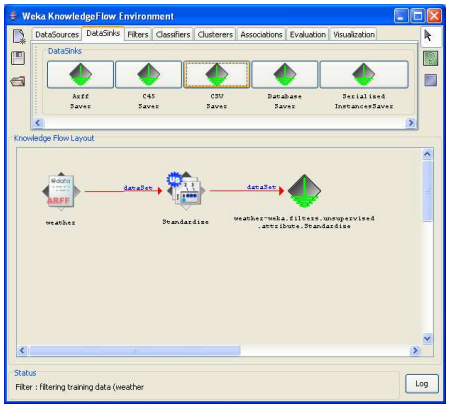
- เริ่มจาก DataSources โดยใช้ ArffLoader
- เลือก Configure… แล้วเลือกแฟ้ม labor.arff
- เลือกแถบ Filters ที่เรียก Normalize เพื่อแปลงตัวแปรที่มีค่าต่อเนื่องให้อยู่ในช่วง [0, 1]
- เลือกแถบ Visualization แล้วเลือก TextViewer เพื่อแสดงผลลัพธ์

การเพิ่มขั้นตอน Standardize



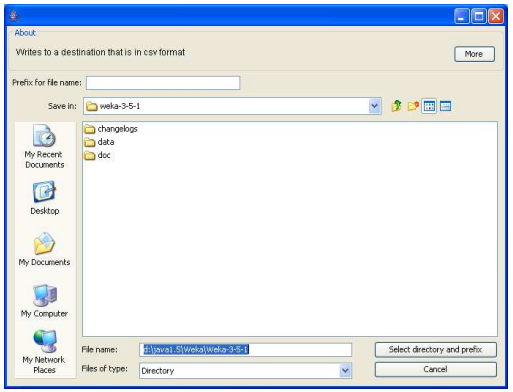
- เริ่มจาก DataSources โดยใช้ ArffLoader
- เลือก Configure… แล้วเลือกแฟ้ม labor.arff
- เลือกแถบ Filters ที่เรียก Standardize เพื่อแปลงตัวแปรให้มีค่าตกอยู่ในช่วงของการกระจายแบบปรกติมาตรฐาน
- เลือกแถบ Visualization แล้วเลือก TextViewer เพื่อแสดงผลลัพธ์

การบันทึกข้อมูลลงแฟ้ม CSV



- ซอฟต์แวร์ Weka สามารถแปลงข้อมูลให้อยู่ในรูปแบบ csv เพื่อนำไปใช้กับซอฟต์แวร์อื่น เช่น Calc
- เริ่มจากเลือก ArffLoader ใน DataSources
- แปลงข้อมูลให้เหมาะสม
- เลือกแถบ DataSinks แล้วเลือก CSVsaver
- เลือกแฟ้มข้อมูล Arff ที่ต้องการ แล้วเลือก Start Loading

การบันทึกลงแฟ้ม CSV ต่อ

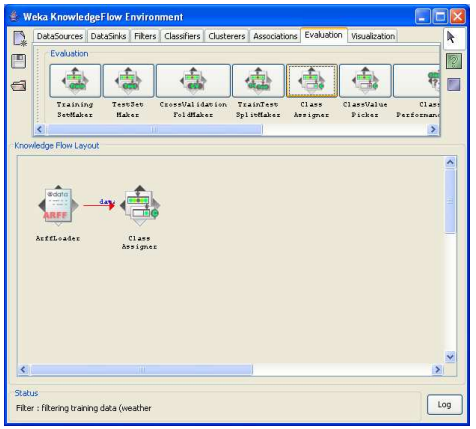


- เลือก Configure… ในเมนูของ CSVsaver
- เปลี่ยนสถานที่ที่ต้องการเก็บไปตำแหน่งที่ต้องการเก็บ โดยเพิ่ม prefix ให้กับชื่อแฟ้มที่ต้องการ
- เก็บข้อมูลโดยเลือก Start loading ใน ArffLoader

โครงการทำเหมืองข้อมูลโดยใช้ผังการไหล

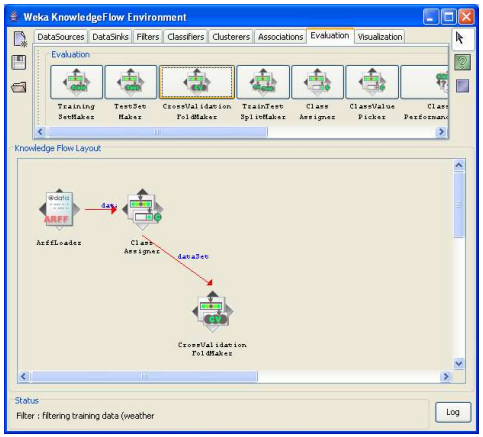
- กำหนดข้อมูลในแฟ้ม iris.arff ให้หาตัวแบบต้นไม้การตัดสินใจที่ดีที่สุด โดยใช้ 5 fold cross-validation กับขั้นตอนวิธี J48 แสดงผลลัพธ์ที่ได้ในรูปแบบต้นไม้
- แนวทางวางผังการไหล: DataSources → Evaluation → J48 → Visualization
  - เริ่มจากการอ่านแฟ้ม iris.arff
  - กำหนดลักษณะประจำที่ใช้แทนคลาส
  - แบ่งข้อมูลออกเป็น 5 ส่วนเพื่อทำ cross-validation
  - ใช้ขั้นตอนวิธี J48
  - แสดงผลลัพธ์

การอ่าน iris.arff



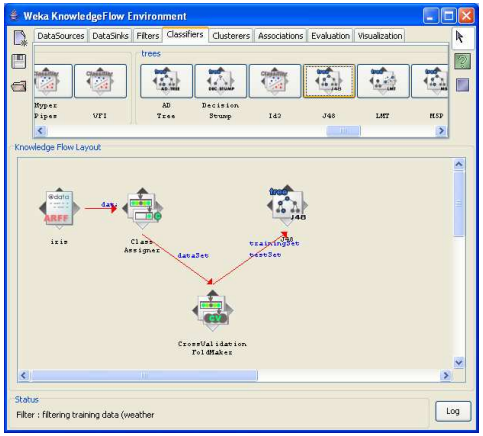
- เลือก ArffLoader จากแถบ DataSources
- ปรับแต่งให้เลือกแฟ้ม iris.arff จาก Configure… เมนู
- เลือก Class Assignment จากแถบ Evaluation
- เลือกคลาสเป้าหมาย

การแยกออกเป็น *k*-fold cross validation



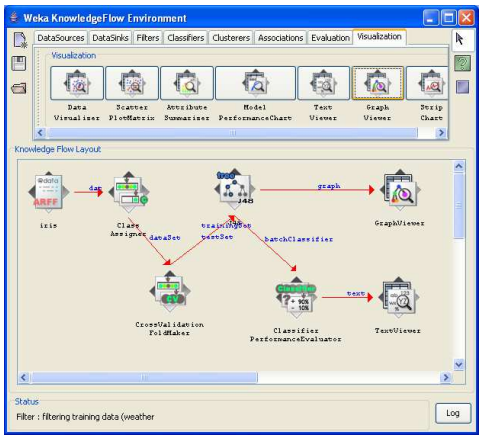
- เลือก CrossValidation FoldMaker จากแถบ Evaluation
- ปรับแต่งให้มีจำนวน fold เท่ากับ 5
- ส่งข้อมูล DataSet จาก Class Assigner

การเรียกใช้ขั้นตอนวิธี J48



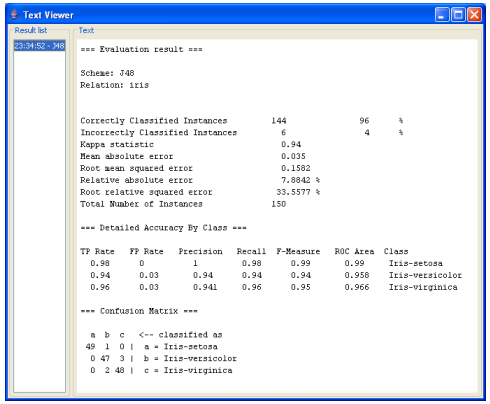
- เลือก J48 จากแถบ Classifiers
- โยงข้อมูลจาก CrossValidation FoldMaker โดยเลือก training set และ test set โยงไปที่ J48
- สั่งให้ข้อมูลนำเข้า โดยเลือก Start loading จาก ArffLoader

การแสดงผลที่ได้จากผังการไหล



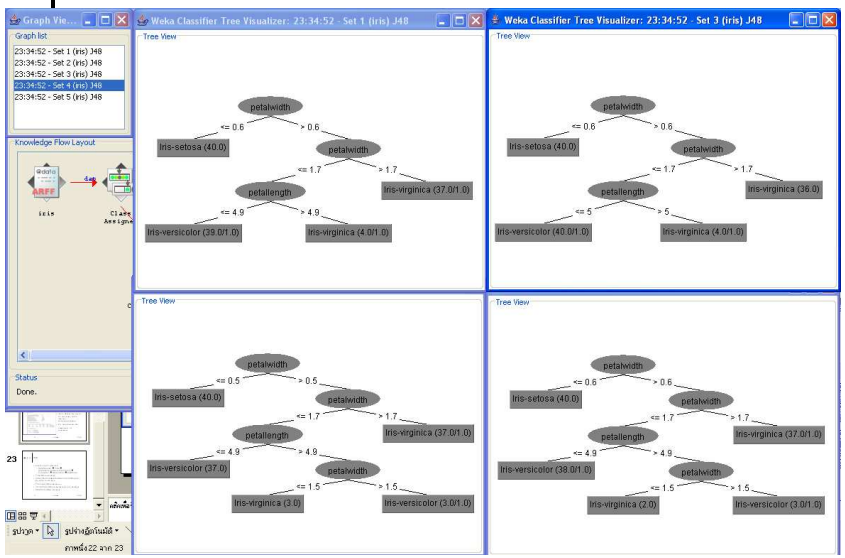
- เลือก Classifiers PerformanceEvaluators จากแถบ Evaluation
- โยงข้อมูลจาก J48 โดยเลือก batchClassifiers โยงไปที่ Classifiers PerformanceEvaluators
- สร้าง TextViewer และ/หรือ GraphViewer จาก Visualization

ผลลัพธ์ที่ได้ในรูปเนื้อความของ J48



- เลือก Start Loading จากแถบ ArffLoader
- รอจนจบ แล้วเลือก Show results จาก TextViewer
- ผลลัพธ์ที่ได้แสดงดังรูปซ้ายซึ่งให้ค่าที่ถูกต้อง 96%
- ใน Confusion Matrix แสดงผลจากการเปรียบเทียบกับกลุ่มที่สนใจ

ผลลัพธ์ในรูปต้นไม้การตัดสินใจ



สรุป

- การออกแบบผังการไหลโดยปรกติ
  - DataSources → Filter → Classifier/Clusterers/Associations → Evaluation → Visualization → DataSinks
- Filter ใช้ในการเตรียมข้อมูล
- Classifier/Clusterers/Associations ใช้ในการสร้างตัวแบบในการทำเหมืองข้อมูล
- Evaluation ใช้ในการเลือกตัวแบบ
- Visualization ใช้ในการแสดงผลลัพธ์ของการทำเหมืองข้อมูล
- DataSinks ใช้ในการเก็บผลลัพธ์