**Lab 8** **204453: Pattern Recognition**

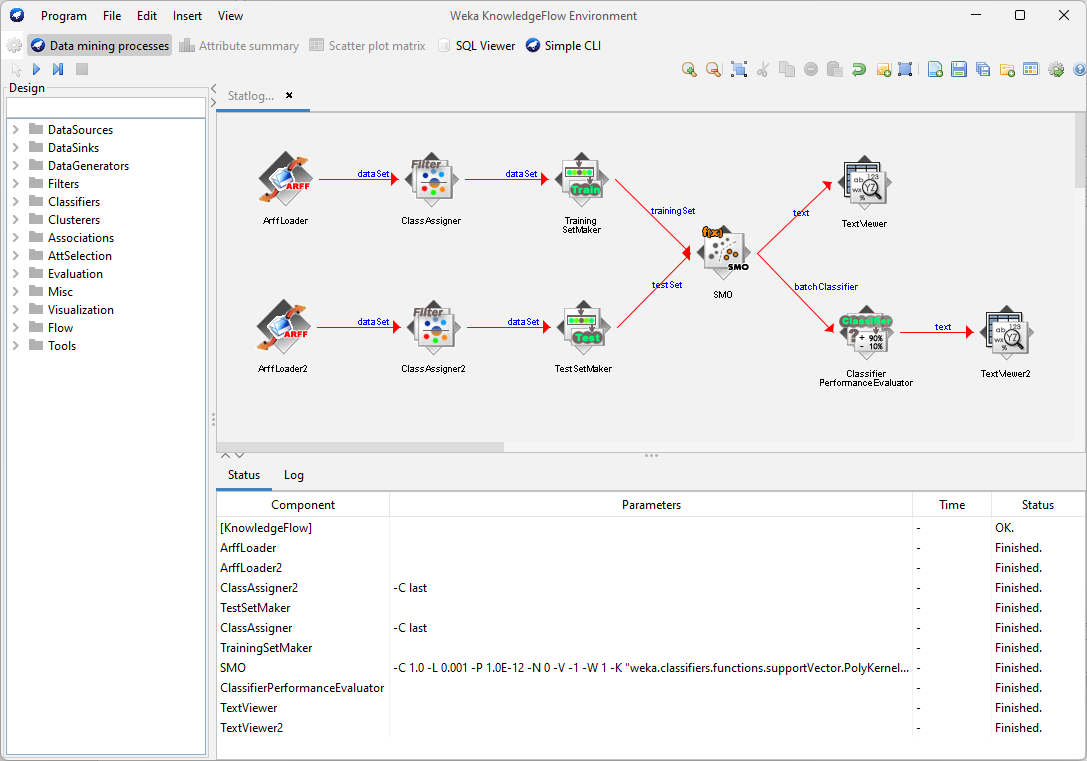
**คำสั่ง:** ให้นักศึกษาดาวน์โหลดไฟล์ชื่อ train.arff และ test.arff จาก Statlog (Landsat Sattelite) จากนั้นให้ทำการทดลอง เครื่องกลเวกเตอร์สนับสนุน (SMO) โดยใช้ Weka Knowledge Flow แล้วบันทึกค่า Linear Discriminant Function จำนวน 1 สมการ

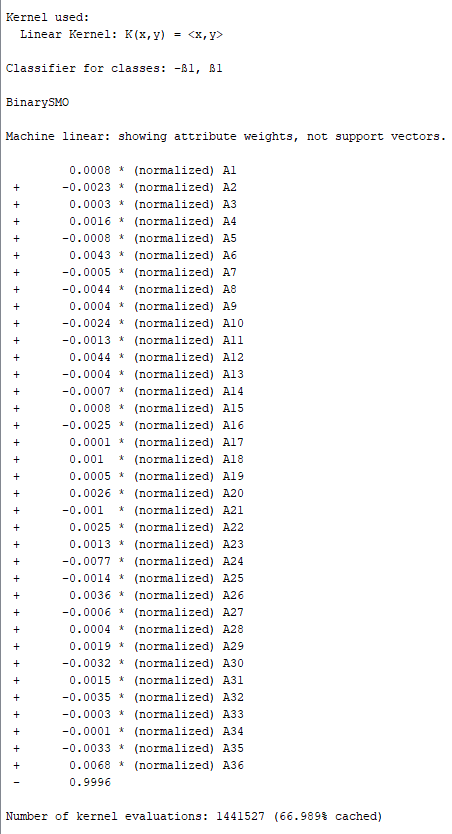
Diagram

Description automatically generated

WEKA Knowledge Flow

1. ไฟล์ Statlog (Landsat Satellite)  
- รายละเอียดการรู้จำแบบ  
Statlog (Landsat Satellite) เป็นไฟล์เกี่ยวกับชุดข้อมูลของรูปภาพถ่ายจากดาวเทียม Landsat ที่ถูกแปลงเป็นตัวเลขแล้วโดยแต่เดิมมีชั้นข้อมูล 7 ชั้นข้อมูลซึ่งระบุลักษณะพื้นที่หรือภูมิประเทศ แต่เราต้องการทำการทดลองชุดข้อมูลนี้ด้วย Support Vector Machine ที่มีชั้นข้อมูลเพียงแค่ 2 ชั้นข้อมูล จึงได้ทำการปรับจำนวนชั้นข้อมูลให้เหลือแค่ 2 ชั้นข้อมูล (Binary Class) เพื่อทำนายชั้นข้อมูล (C) ว่าข้อมูลจะอยู่ในชั้นข้อมูล -1 หรือชั้นข้อมูล 1  
- ชั้นข้อมูล  
จำนวนชั้นข้อมูลถูกลดลงจาก 7 ชั้นข้อมูลเหลือ 2 ชั้นข้อมูล จึงเป็น Binary Class คือ -1 และ 1  
- ค่าทางสถิติ  
ชุดข้อมูล Statlog (Landsat Satellite) ถูกแบ่งเป็น 2 ชุดข้อมูลดังนี้  
 - ชุดข้อมูลแรกเป็นชุดข้อมูลสำหรับการ Train มีจำนวนระเบียน 4435 ระเบียน, มีจำนวนสดมภ์ 37 สดมภ์, มีจำนวนชั้นข้อมูล 2 ชั้นข้อมูล และ จำนวนข้อมูลในแต่ละชั้นข้อมูล คือ ชั้นข้อมูล -1 มีจำนวนข้อมูล 4020 ข้อมูล และชั้นข้อมูล 1 มีจำนวนข้อมูล 415 ข้อมูล  
 - ชุดข้อมูลที่สองเป็นชุดข้อมูลสำหรับการ Test มีจำนวนระเบียน 2000 ระเบียน, มีจำนวนสดมภ์ 37 สดมภ์, มีจำนวนชั้นข้อมูล 2 ชั้นข้อมูล และ จำนวนข้อมูลในแต่ละชั้นข้อมูล คือ ชั้นข้อมูล -1 มีจำนวนข้อมูล 1789 ข้อมูล และชั้นข้อมูล 1 มีจำนวนข้อมูล 211 ข้อมูล

2. แสดงการทำงานของ Weka Knowledge Flow   


3. ผลการทดลองของเครื่องกลเวกเตอร์สนับสนุน (Support Vector Machine: SVM)   
- บันทึกค่า Linear Discriminant Function จำนวน 1 สมการได้ดังนี้  


- Model Accuracy by Confusion Matrix

**SMO** usingParameters: (batchSize=100, c = 1.0, calibrator = Logistic -R 1.0E-8 -M -1 -num-decimal-places 4)

|  |  |  |  |
| --- | --- | --- | --- |
|  | *true positive* | *true negative* | *class precision* |
| *Pred. positive* | 0 | 0 | ? |
| *Pred. negative* | 211 | 1789 | 89.50% |
| *class recall* | 0.00% | 100.00% |  |

4. สรุปผลการทดลอง  
ความแม่นยำในการทำนายภาพถ่ายทางดาวเทียมว่าอยู่ในชั้นข้อมูล -1 นั้นสูงมาก แต่ความแม่นยำในการทำนายว่าข้อมูลจะอยู่ในชั้นข้อมูล 1 เป็น 0.00% กล่าวคือ ไม่มีความแม่นยำเลย ดังนั้นค่าความแม่นยำโดยรวมของชุดข้อมูล Statlog (Landsat Satellite) ที่เป็น 89.45% นั้นมีความ Bias สูงมาก เนื่องจากชั้นข้อมูล -1 มีจำนวน 1789 ระเบียน แต่ชั้นข้อมูล 1 มีจำนวนข้อมูลแค่ 211 ระเบียนเท่านั้นเอง การจะลดความ Bias นี้ลงได้อาจจะต้องใช้ตัวแบบการจำแนกอื่นที่เหมาะสมกับการทำนายชุดข้อมูลนี้ เช่น Random Forest, LightGBM เป็นต้น นอกจากนี้ความเป็น Class Imbalance ของชุดข้อมูลนี้ ยังทำให้ต้องทำให้ทั้งสองชั้นข้อมูลมีจำนวนข้อมูลที่ใกล้เคียงกันด้วยวิธีต่าง ๆ เช่น การสังเคราะห์ข้อมูลด้วย SMOTE, การ Duplicate ฯลฯ

URL: <https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)>

**หมายเหตุ:** รายงาน ให้ใช้รูปแบบของ Lab 5