CSC662 ปฏิบัติการ ๖

ตัวกรองที่ใช้ในซอฟต์แวร์ Weka

เขียนโดย ผศ. ดร. กรุง สินอภิรมย์สราญ

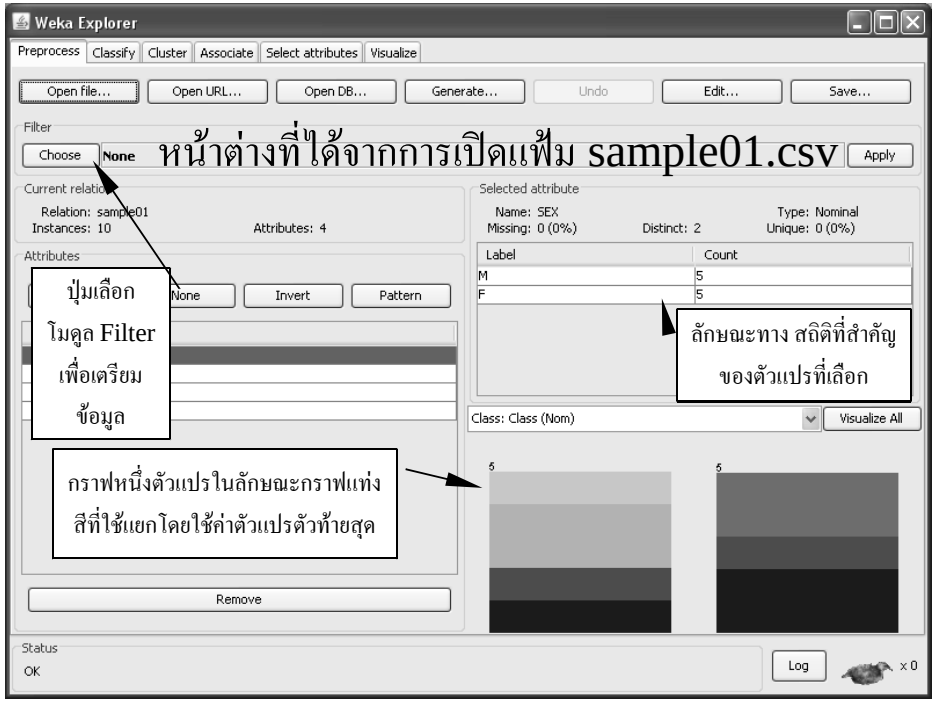
ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

เนื้อหาที่ครอบคลุม

- ตัวกรองในซอฟต์แวร์ Weka
- การกำจัดลักษณะประจำ
- ตัวกรองแบบอัตโนมัติ Supervised filter
  - ลักษณะประจำ
  - ระเบียน
- ตัวกรองที่ผู้ใช้กำหนดเอง Unsupervised filter
  - ลักษณะประจำ
  - ระเบียน

ตัวอย่างแฟ้มข้อมูล sample01.csv

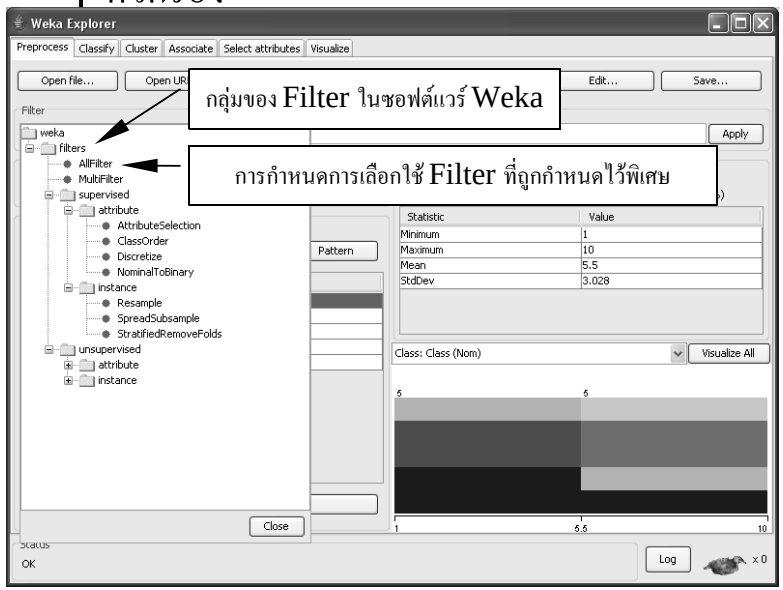
```
ID,SEX,PASS/FAIL,Score,Class
1,M,Pass,45.5,B
2,F,Pass,56.78,B
3,M,Pass,89,A
4,F,Pass,77,A
5,M,Fail,32,C
6,F,Fail,12,D
7,M,Fail,35,C
8,F,Pass,62,B
9,M,Pass,68,B+
10,F,Fail,10,D
```

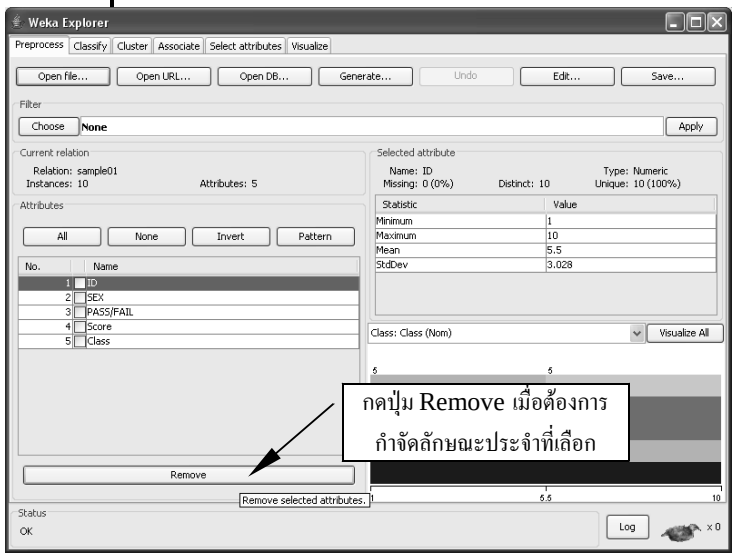
การใช้ตัวกรองในซอฟต์แวร์ Weka

- ตัวกรอง (Filters) รวบรวมโมดูลในขั้นการเตรียมข้อมูล
- ตัวกรองแบ่งออกเป็นสองลักษณะคือ
  - Supervised รวมโมดูลที่แปลงข้อมูลแบบอัตโนมัติที่มีการควบคุมด้วยพารามิเตอร์ที่ผู้ใช้กำหนด แบ่งเป็นสองหมวดใหญ่ คือลักษณะประจำ (attribute) กับข้อมูลแต่ละระเบียน (instance)
  - Unsupervised รวมโมดูลที่แปลงข้อมูลที่ผู้ใช้กำหนดเอง แบ่งเป็นสองหมวดใหญ่คือ ลักษณะประจำ (attribute) กับข้อมูลแต่ละระเบียน (instance)

ตัวกรอง



การกำจัดลักษณะประจำ

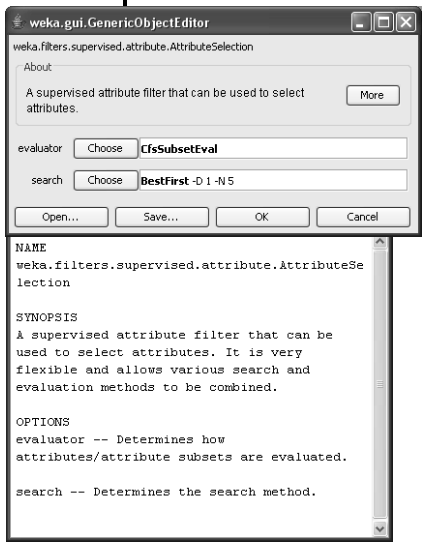


- เราสามารถเลือกลักษณะประจำที่ไม่ต้องการออกได้โดยการเลือกลักษณะประจำที่ต้องการ แล้วกดปุ่ม Remove

ตัวกรองแบบอัตโนมัติ Supervised

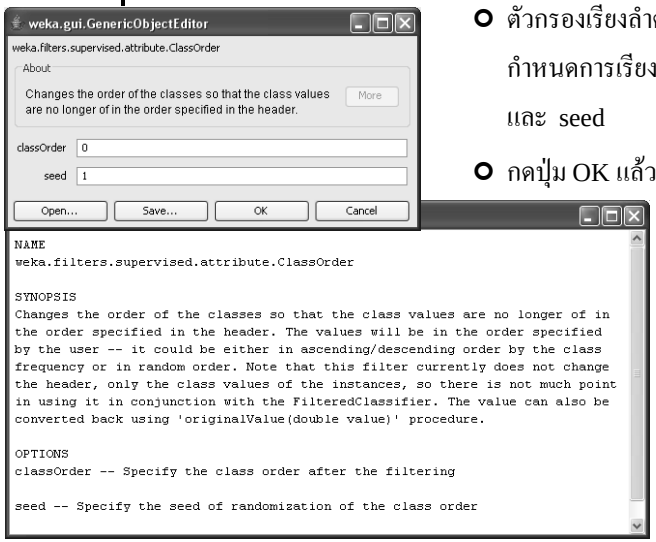
- ประกอบด้วย
  - ลักษณะประจำ: AttributeSelection, ClassOrder, Discretize, NominalToBinary
  - ระเบียน: Resample, SpreadSubsample, StratifiedRemoveFolds

AttributeSelection



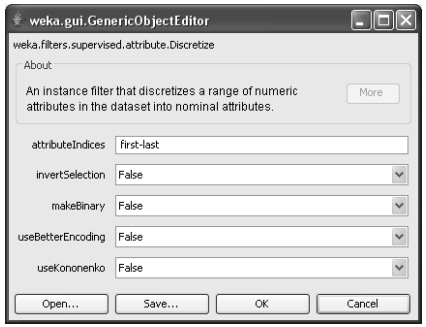
- ตัวกรองที่เลือกลักษณะประจำที่นำมาวิเคราะห์แบบอัตโนมัติ โดยผู้ใช้กำหนดตัวประเมินในกล่อง evaluator และวิธีการค้นในกล่อง search
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

ClassOrder



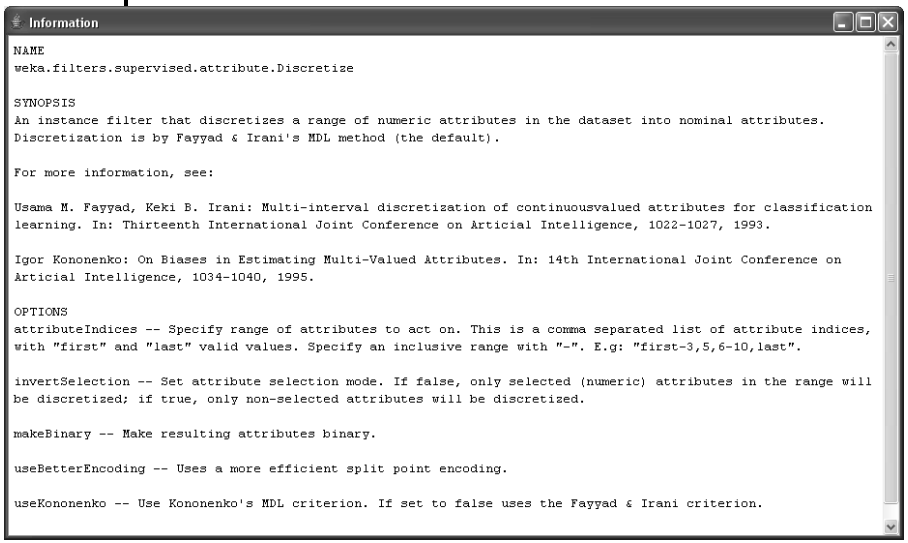
- ตัวกรองเรียงลำดับคลาส โดยผู้ใช้กำหนดการเรียงในกล่อง classOrder และ seed
- กดปุ่ม OK แล้วกดปุ่ม Apply

Discretize

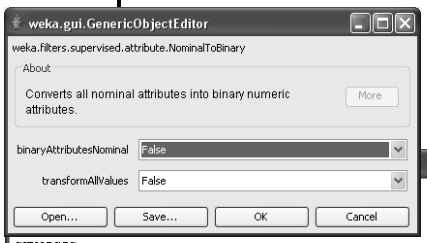


- ตัวกรองแปลงค่าต่อเนื่องให้เป็นค่าไม่ต่อเนื่อง โดยผู้ใช้เลือกลักษณะประจำที่ต้องการเปลี่ยนในกล่อง attributeIndices และผู้ใช้กำหนดตัวเลือกโดยดู Help ในหน้าถัดไป
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

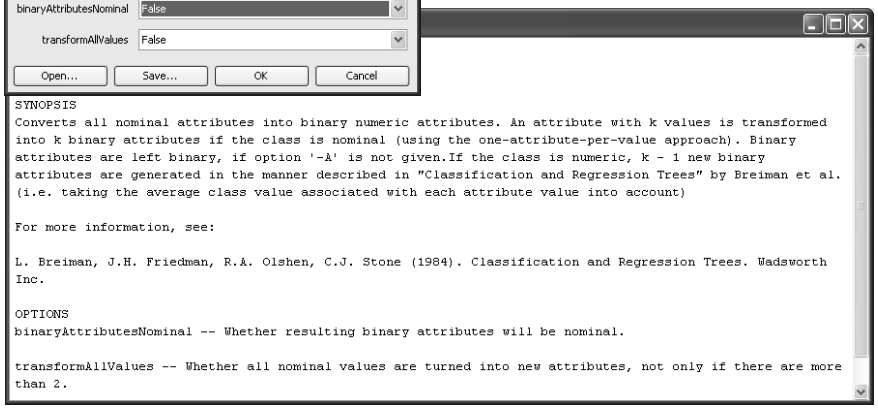
Discretize Help



NominalToBinary

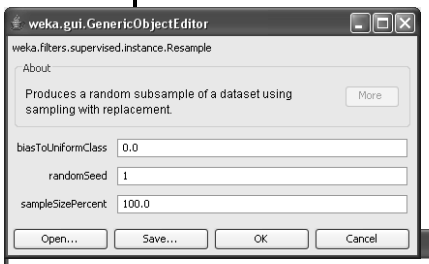


- ตัวกรองแปลงค่า Nominal ให้เป็นค่า Binary โดยผู้ใช้กำหนดค่าตัวเลือก
- กดปุ่ม OK แล้วกดปุ่ม Apply

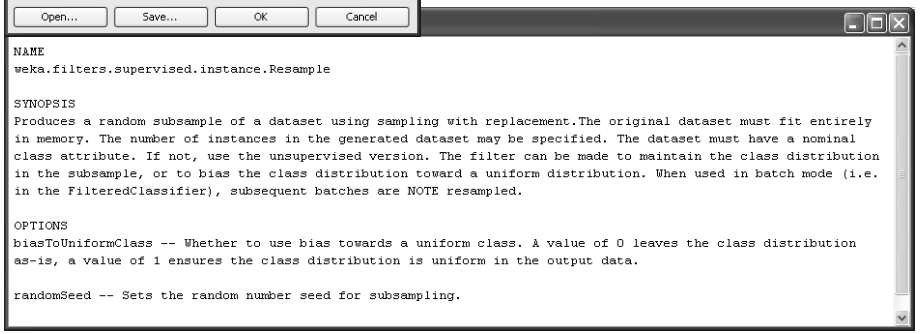


```
SYNOPSIS
Converts all nominal attributes into binary numeric attributes. An attribute with k values is transformed
into k binary attributes if the class is nominal (using the one-attribute-per-value approach). Binary
attributes are left binary, if option '-A' is not given.If the class is numeric, k - 1 new binary
attributes are generated in the manner described in "Classification and Regression Trees" by Breiman et al.
(i.e. taking the average class value associated with each attribute value into account)

For more information, see:

L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone (1984). Classification and Regression Trees. Wadsworth
Inc.

OPTIONS
binaryAttributesNominal -- Whether resulting binary attributes will be nominal.

transformAllValues -- Whether all nominal values are turned into new attributes, not only if there are more
than 2.
```

Resample

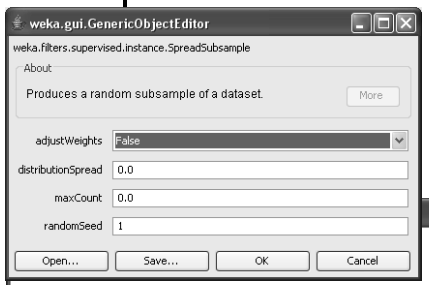


- ตัวกรองที่ใช้สุ่มเลือกตัวอย่าง โดยมีการสุ่มแบบแทนค่ากลับคืน โดยผู้ใช้กำหนดค่าจำนวนตัวอย่างในกล่อง sampleSizePercent
- กดปุ่ม OK แล้วกดปุ่ม Apply

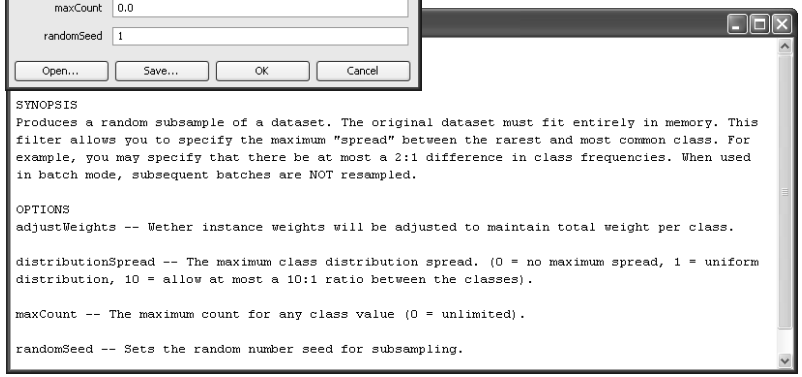


```
NAME
weka.filters.supervised.instance.Resample

SYNOPSIS
Produces a random subsample of a dataset using sampling with replacement.The original dataset must fit entirely
in memory. The number of instances in the generated dataset may be specified. The dataset must have a nominal
class attribute. If not, use the unsupervised version. The filter can be made to maintain the class distribution
in the subsample, or to bias the class distribution toward a uniform distribution. When used in batch mode (i.e.
in the FilteredClassifier), subsequent batches are NOTE resampled.

OPTIONS
biasToUniformClass -- Whether to use bias towards a uniform class. A value of 0 leaves the class distribution
as-is, a value of 1 ensures the class distribution is uniform in the output data.

randomSeed -- Sets the random number seed for subsampling.
```

SpreadSubsample

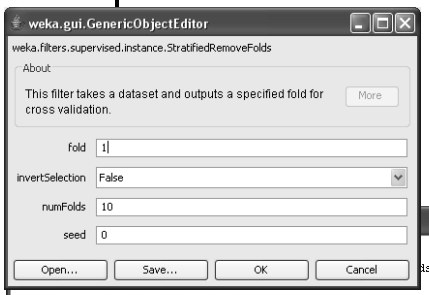


- ตัวกรองที่ใช้สุ่มเลือกตัวอย่างที่มีการกำหนดสัดส่วนของการกระจาย โดยผู้ใช้กำหนดค่าการกระจายในกล่อง distributionSpread
- กดปุ่ม OK แล้วกดปุ่ม Apply

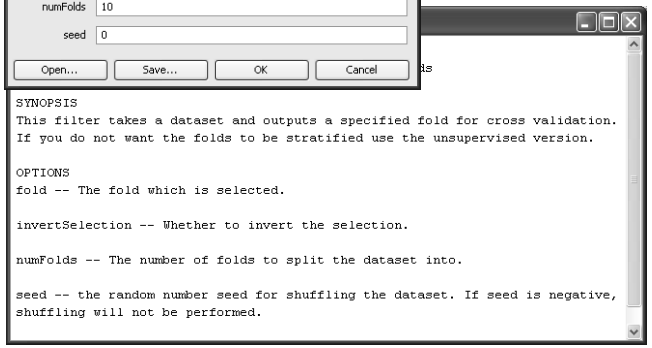


```
SYNOPSIS
Produces a random subsample of a dataset. The original dataset must fit entirely in memory. This
filter allows you to specify the maximum "spread" between the rarest and most common class. For
example, you may specify that there be at most a 2:1 difference in class frequencies. When used
in batch mode, subsequent batches are NOT resampled.

OPTIONS
adjustWeights -- Wether instance weights will be adjusted to maintain total weight per class.

distributionSpread -- The maximum class distribution spread. (0 = no maximum spread, 1 = uniform
distribution, 10 = allow at most a 10:1 ratio between the classes).

maxCount -- The maximum count for any class value (0 = unlimited).

randomSeed -- Sets the random number seed for subsampling.
```

StratifiedRemoveFolds



- ตัวกรองที่ใช้สุ่มเลือกกลุ่มตัวอย่างออกเป็นชุด fold โดยผู้ใช้กำหนดตัวเลือก
- กดปุ่ม OK
- แล้วกดปุ่ม Apply



```
SYNOPSIS
This filter takes a dataset and outputs a specified fold for cross validation.
If you do not want the folds to be stratified use the unsupervised version.

OPTIONS
fold -- The fold which is selected.

invertSelection -- Whether to invert the selection.

numFolds -- The number of folds to split the dataset into.

seed -- the random number seed for shuffling the dataset. If seed is negative,
shuffling will not be performed.
```

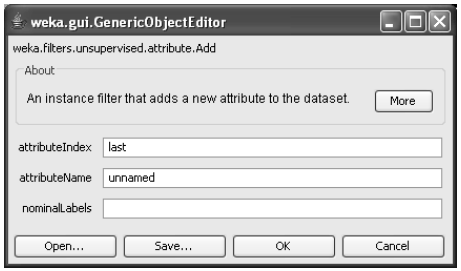
ตัวกรองที่ผู้ใช้กำหนดเอง Unsupervised

- เราจะเลือกอธิบายตัวกรองบางตัวเท่านั้น สำหรับตัวกรองอื่น ผู้ใช้สามารถอ่านได้จาก Help ของซอฟต์แวร์ Weka
- ลักษณะประจำ: Add, AddCluster, AddExpression, AddNoise, ClusterMembership, Copy, Discretize, FirstOrder, MakeIndicator, MergeTwoValues, NominalToBinary, Normalize, NumericToBinary, NumericTransform, Obfuscate, PKIDiscretize, RandomProjection, Remove, RemoveType, RemoveUseless, ReplaceMissingValues, Standardize, StringToNominal, StringToWordVector, SwapValues, TimeSeriesDelta, TimeSeriesTranslate
- ระเบียน: Normalize, NonSparseToSpare, Randomize, RemoveFolds, RemoveMisclassified, RemovePercentage, RemoveRange, RemoveWithValues, Resample, SparseToNonSparse

ตัวกรองที่ผู้ใช้กำหนดเองกับลักษณะประจำ

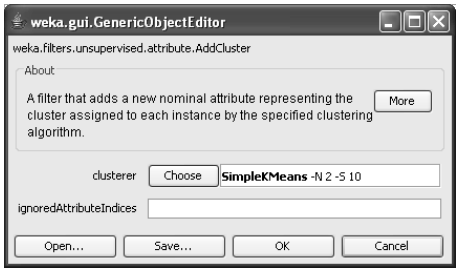
- Add filter
- AddExpression filter
- NominalToBinary filter
- NumericToBinary filter
- NumericTransform filter
- Remove filter
- ReplaceMissingValues filter
- Standardize filter
- AddCluster filter
- Discretize filter
- Normalize filter
- RemoveType filter

Add filter



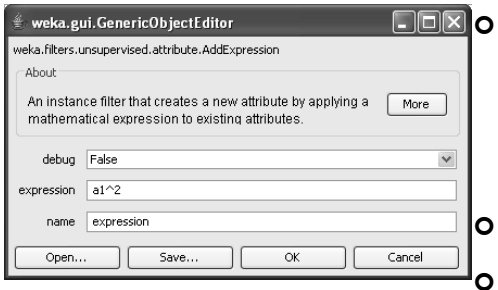
- ตัวกรองเพิ่มลักษณะประจำ เลือก add โดยเพิ่มลักษณะประจำที่มีค่าตั้งต้นคือ missing value
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

AddCluster filter



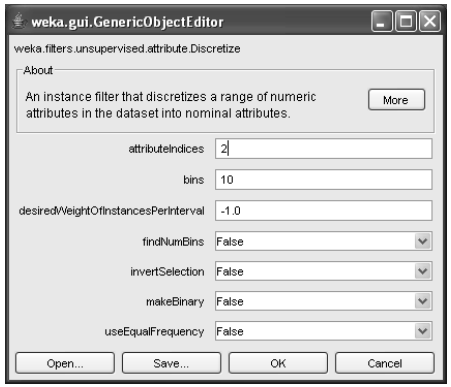
- ตัวกรองเพิ่มลักษณะประจำตามการเกาะกลุ่ม เลือก addCluster เลือกวิธีการเกาะกลุ่ม เช่น SimpleKMeans
- กำหนดลักษณะประจำที่ไม่นำมาใช้ในการวิเคราะห์การเกาะกลุ่มใน ignoredAttributeIndices
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

AddExpression filter



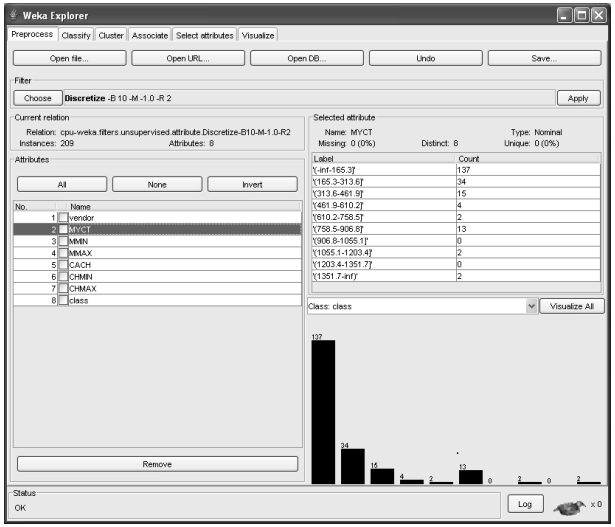
- ตัวกรองเพิ่มลักษณะประจำตามนิพจน์จากลักษณะประจำที่กำหนด เลือก addExpression พิมพ์นิพจน์ที่ต้องการสร้างลักษณะประจำใหม่
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

Discretize filter

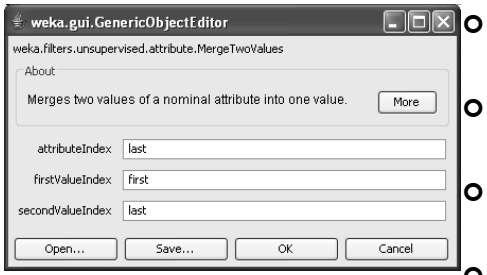


- ตัวกรองแปลงเป็นค่าไม่ต่อเนื่อง ผู้ใช้เลือกลักษณะประจำในช่อง attributeIndices ตามลำดับลักษณะประจำที่กำหนด
- กำหนดจำนวนกล่องที่ต้องการใน bins
- เราสามารถแบ่งแบบ equal width หรือ equal depth โดยปรับเป็น False ที่ useEqualFrequency
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

ผลการใช้ตัวกรอง Discretize

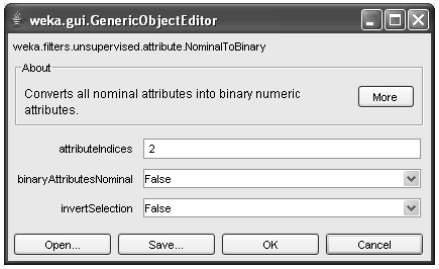


MergeTwoValues filter



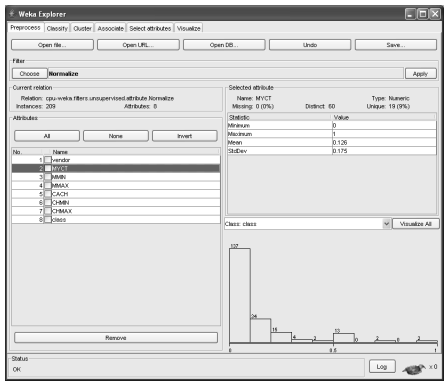
- ตัวกรองรวมค่าสองค่าเป็นหนึ่ง เลือก MergeTwoValues
- กำหนดดรรชนีของลักษณะประจำใน attributeIndex
- กำหนดค่าใน firstValueIndex และ secondValueIndex
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

NominalToBinary filter



- เลือกตัวกรองแปลงค่าไม่ต่อเนื่องเป็นค่า 0 หรือ 1 เลือก NominalToBinary
- กำหนดดรรชนีของลักษณะประจำใน attributeIndices ที่ต้องการ
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

Normalize filter

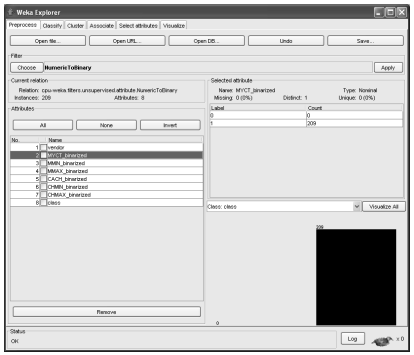


- ตัวกรองเปลี่ยนเป็นค่ามาตรฐานเลือก Normalize เพื่อปรับลักษณะประจำทุกลักษณะประจำเฉพาะลักษณะประจำที่เป็นจำนวน จะถูกแปลงให้มีค่าอยู่ในช่วง 0-1 โดยใช้สูตร

$$v'_i = \frac{v_i - min_A}{max_A - min_A}$$

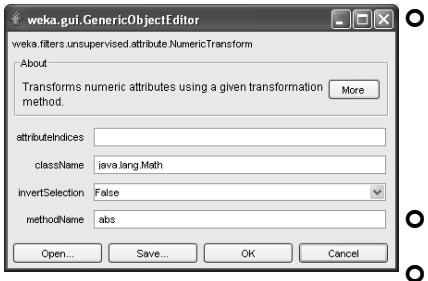
- กดปุ่ม Apply

NumericToBinary filter



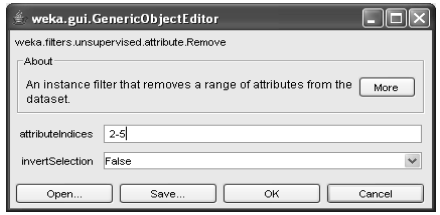
- ตัวกรองแปลงข้อมูลจำนวนให้เป็นค่า 0 หรือ 1 เลือก NumericToBinary โดยเปลี่ยนทุกลักษณะประจำที่เป็นจำนวน ค่าจำนวนที่เป็น 0 จะยังคงค่า 0 แต่ค่าที่ไม่ใช่ 0 จะเปลี่ยนเป็น 1 ทั้งหมด
- กดปุ่ม Apply

NumericTransform filter



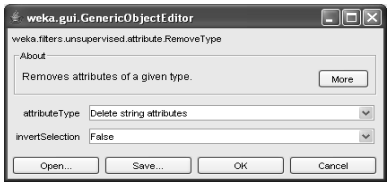
- ตัวกรองแปลงโดยใช้ฟังก์ชันจำนวน เลือก NumericTransform จะแปลงค่าในลักษณะประจำตามฟังก์ชันที่กำหนด เช่น abs
- กดปุ่ม OK
- แล้วกด Apply

Remove filter



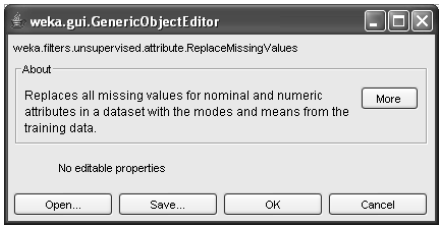
- ตัวกรองกำจัดลักษณะประจำ เลือก Remove โดย attributeIndices
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

RemoveType filter



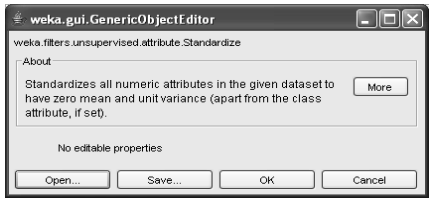
- ตัวกรองกำจัดลักษณะประจำตามชนิดของลักษณะประจำ เลือก RemoveType โดยเลือกชนิดที่ต้องการกำจัดใน attributeType
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

ReplaceMissingValue



- ตัวกรองการแทนค่าที่ขาดหายไปเลือก ReplaceMissingValue
  - แทนด้วยค่าเฉลี่ยสำหรับลักษณะประจำที่เป็นจำนวน
  - แทนด้วยฐานนิยมสำหรับลักษณะประจำที่เป็นค่าไม่ต่อเนื่อง
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

Standardize filter



- ตัวกรองเปลี่ยนข้อมูลให้อยู่ในรูปที่มีการแจกแจงมาตรฐานโดยใช้ z-score โดยเลือก Standardize

$$v'_i = \frac{v_i - \bar{V}}{std_A}, \bar{V} = \sum_{i=1}^{n} \frac{v_i}{n}$$

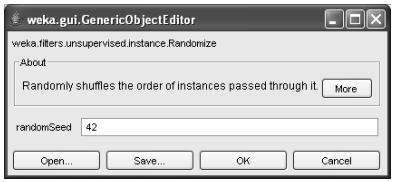
$$std_A = \sum_{i=1}^{n} \frac{(v_i - \bar{V})^2}{n-1}$$

- กดปุ่ม OK
- แล้วกดปุ่ม Apply

ตัวกรองที่ตัวกรองที่ผู้ใช้กำหนดเองกับระเบียน

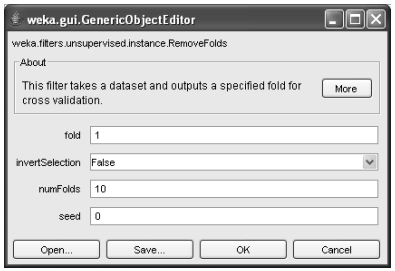
- Randomize
- RemoveFolds
- RemovePercentage
- RemoveRange
- RemoveWithValues
- Resample

Randomize filter



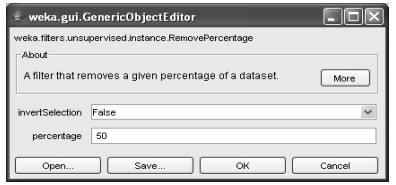
- ตัวกรองสลับสุ่ม เลือก Randomize เพื่อให้มีการเรียงระเบียนแบบสุ่ม
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

RemoveFold filter



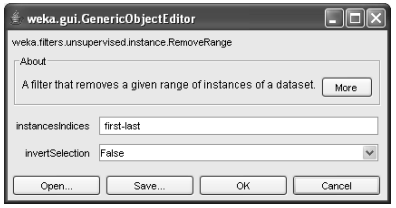
- ตัวกรองกำจัดชุดระเบียน เลือก RemoveFold เพื่อกำจัดข้อมูลตามจำนวนชุด ตามจำนวนชุดทั้งหมดใน numFolds
- กดปุ่ม Save เพื่อบันทึกชุดระเบียน
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

RemovePercentage filter



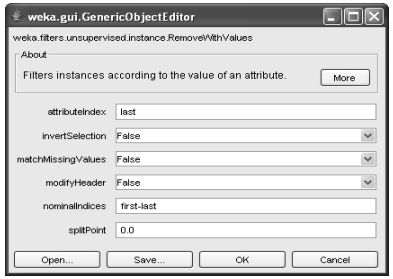
- ตัวกรองกำจัดระเบียนตามเปอร์เซ็นต์ เลือก RemovePercentage เพื่อลดจำนวนข้อมูลโดยเอาออกเท่ากับจำนวนเปอร์เซ็นต์ที่กำหนดใน percentage
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

RemoveRange filter



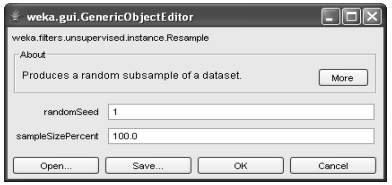
- ตัวกรองกำจัดระเบียนในพิสัยที่กำหนด เลือก RemoveRange เพื่อลดจำนวนข้อมูลที่กำหนดใน instancesindices
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

RemoveWithValues filter



- ตัวกรองกำจัดข้อมูลตามค่า เลือก RemoveWithValues เพื่อลดจำนวนข้อมูลออกโดยใช้ attributeIndex
- ค่าที่ต่ำกว่า splitPoint จะถูกกำจัดทิ้ง
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

Resample filter



- ตัวกรองสุ่มใหม่ เลือก Resample เพื่อให้มีการสุ่มข้อมูลใหม่ โดยกำหนดเป็นเปอร์เซ็นต์ใน sampleSizePercent
- กดปุ่ม save เพื่อบันทึกข้อมูล
- กดปุ่ม OK
- แล้วกดปุ่ม Apply

สรุป

- โมดูลในการเตรียมข้อมูลในซอฟต์แวร์ Weka เรียก ตัวกรอง (Filters) แบ่งออกเป็น
  - Supervised
  - Unsupervised
- นอกจากนี้เราเลือกใช้ตัวกรองกับลักษณะประจำ หรือระเบียน