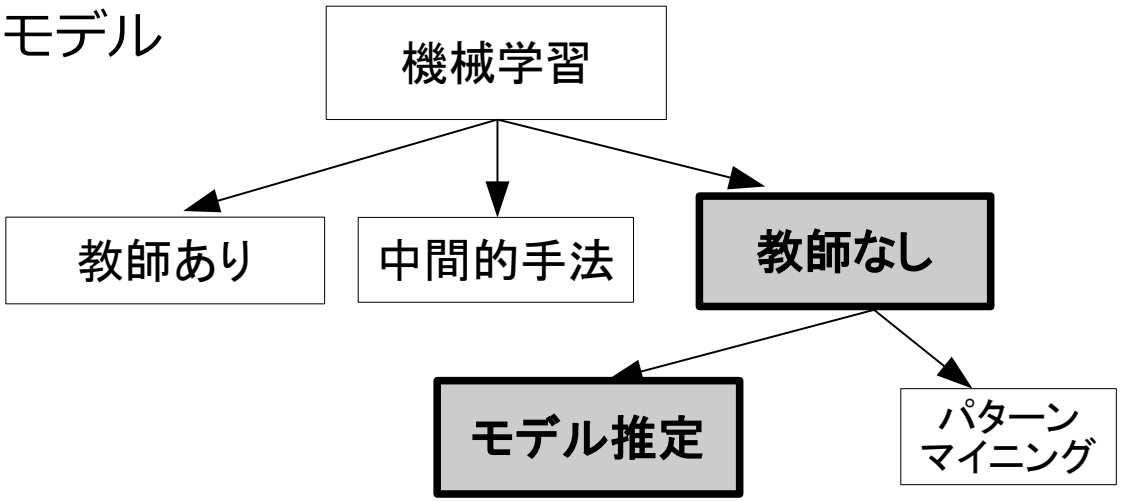
10. モデル推定

- 問題設定
  - 教師なし学習
  - 数値入力　→　クラスモデル
    - クラスモデルの例
      - クラスの分割結果
      - クラスの確率モデル

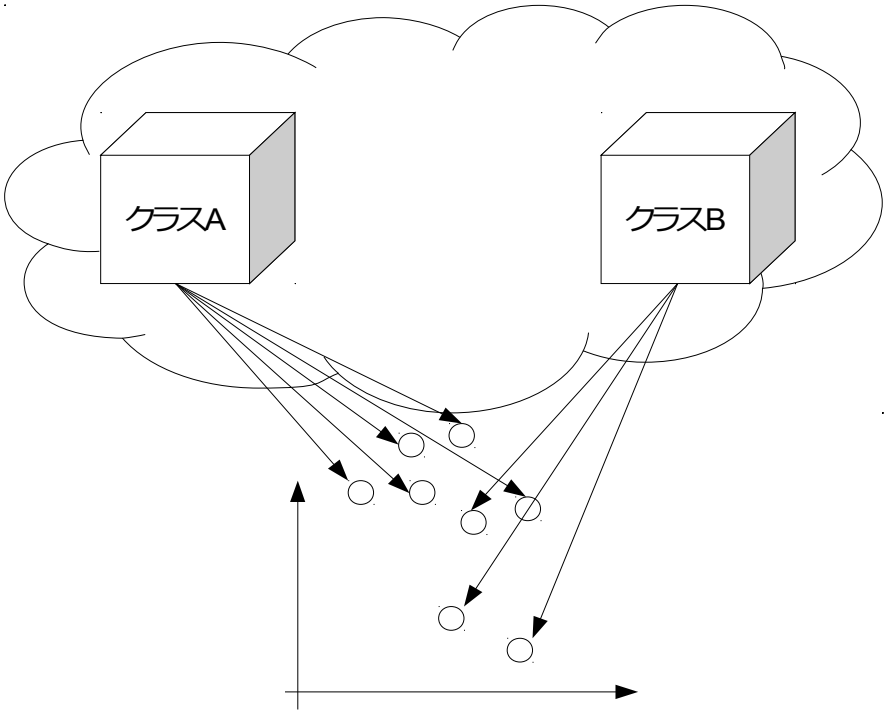


10.1 問題の定義

- 学習データ

$$\{\boldsymbol{x}^{(i)}\} \quad i = 1, .., N$$

- 問題設定
  - 特徴ベクトル $\boldsymbol{x}$ が生成された元のクラスの性質を推定する



10.2 クラスタリング

- クラスタリングとは
  - 対象のデータを、

    内的結合（同じ集合内のデータ間の距離は小さく）

    と

    外的分離（異なる集合間の距離は大きく）

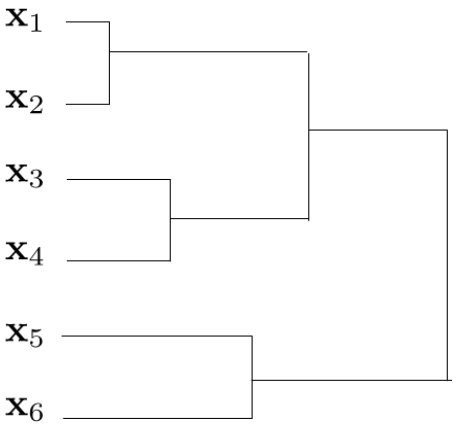
    が達成されるような部分集合に分割すること

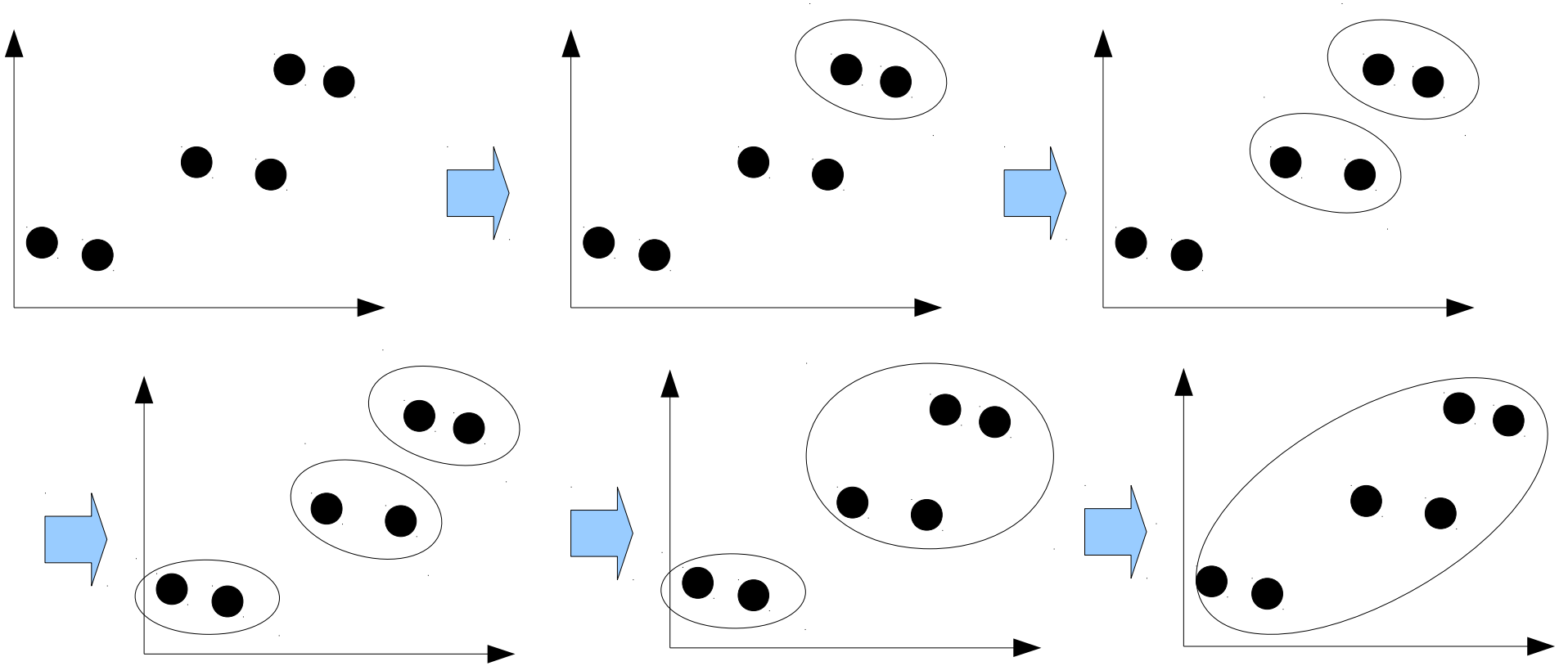
要するに塊を見つけること

- クラスタリング手法の分類
  - 階層的手法
    - ボトムアップ的にデータをまとめてゆく
  - 分割最適化手法
    - トップダウン的にデータ集合を分割してゆく

10.2.1 階層的クラスタリング

- 階層的クラスタリングとは
  1.1 データ 1 クラスタからスタート
  2.最も近接するクラスタをまとめる
  3.全データが 1 クラスタになれば終了





10.2.2 分割最適化クラスタリング ― k-means アルゴリズム―

- 分割最適化クラスタリングとは
  - データ分割の良さを評価する関数を定め、その評価関数の値を最適化することを目的とする
  - ただし、全ての可能な分割に対して評価値を求めることは、データ数 N が大きくなると、不可能
    - 2 分割で 2^N 通り
  - 探索によって、準最適解を求める

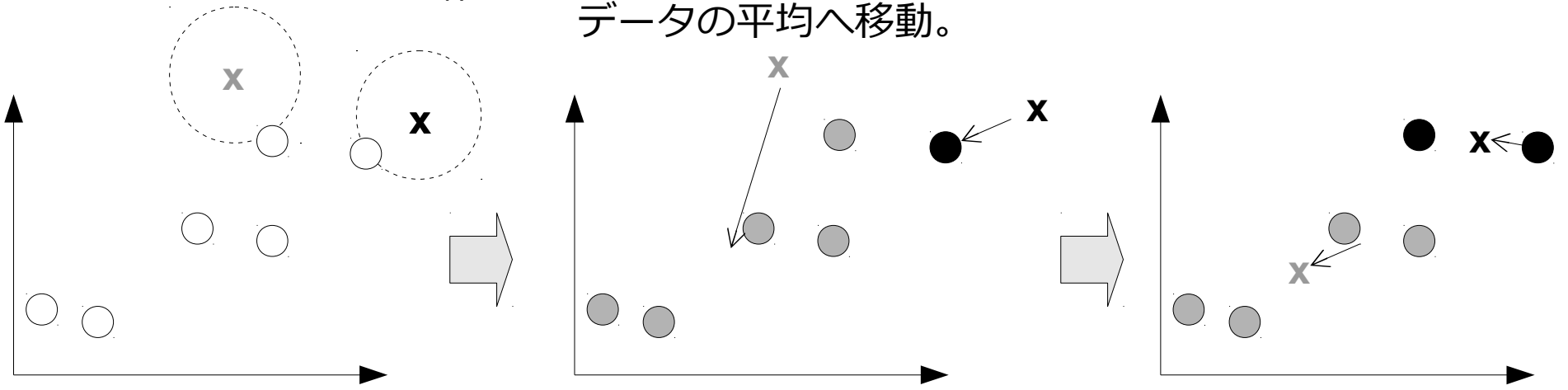
10.2.2 分割最適化クラスタリング ― k-means アルゴリズム―

- k-Means アルゴリズム
  1. 分割数 k を予め与える
  2. 乱数で k 個のクラスタ中心を設定し、逐次更新

k=2 とし、初期値として
乱数でクラスタ中心を配置

全データを近い方のクラスタ
中心に所属させる。そして、
クラスタ中心を所属している
データの平均へ移動。

左の処理を繰り返す。



10.2.3 自動分割最適化クラスタリング ― X-means アルゴリズム―

- k-means 法の問題点
  - 分割数 k を予め決めなければならない
- 解決法 ⇒ X-means アルゴリズム
  - 2 分割から始めて、分割数を適応的に決定する
  - 分割の妥当性の判断： BIC(Bayesian information criterion) が小さくなれば、分割を継続

$$BIC = -2\log L + q\log N$$

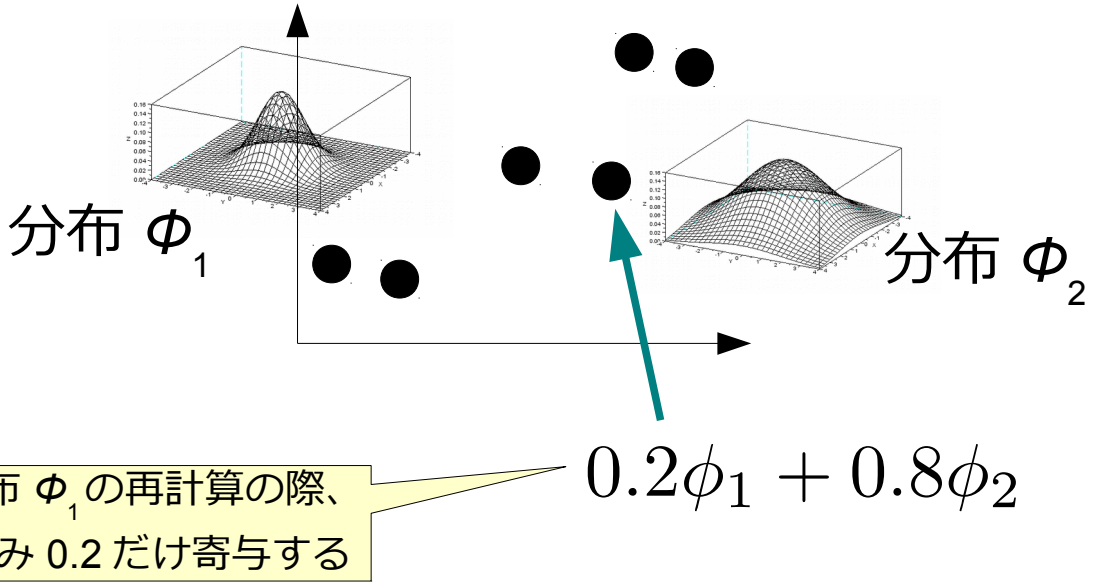
  - L: モデルの尤度
  - q: モデルのパラメータ数
  - N: データ数

パラメータで表される
統計モデルの選択基準
(小さいほどよいモデル)

10.4 確率密度推定

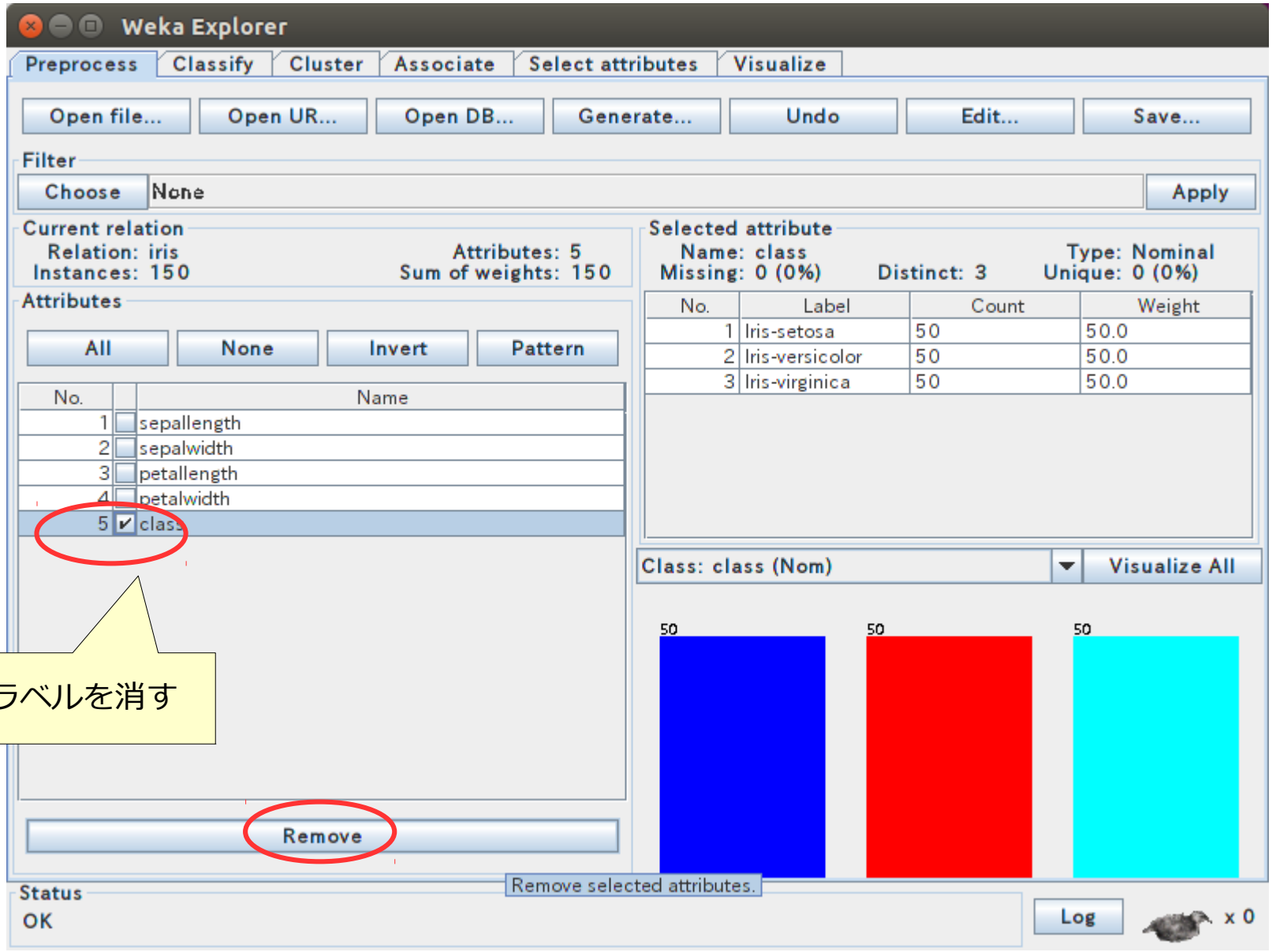
- 教師なし学習で識別器を作る問題
  - クラスタリング結果からは、 1 クラス 1 プロトタイプの単純な識別器しかできない
  - 各クラスの事前確率や確率密度関数も推定したい

⇨ EM アルゴリズム



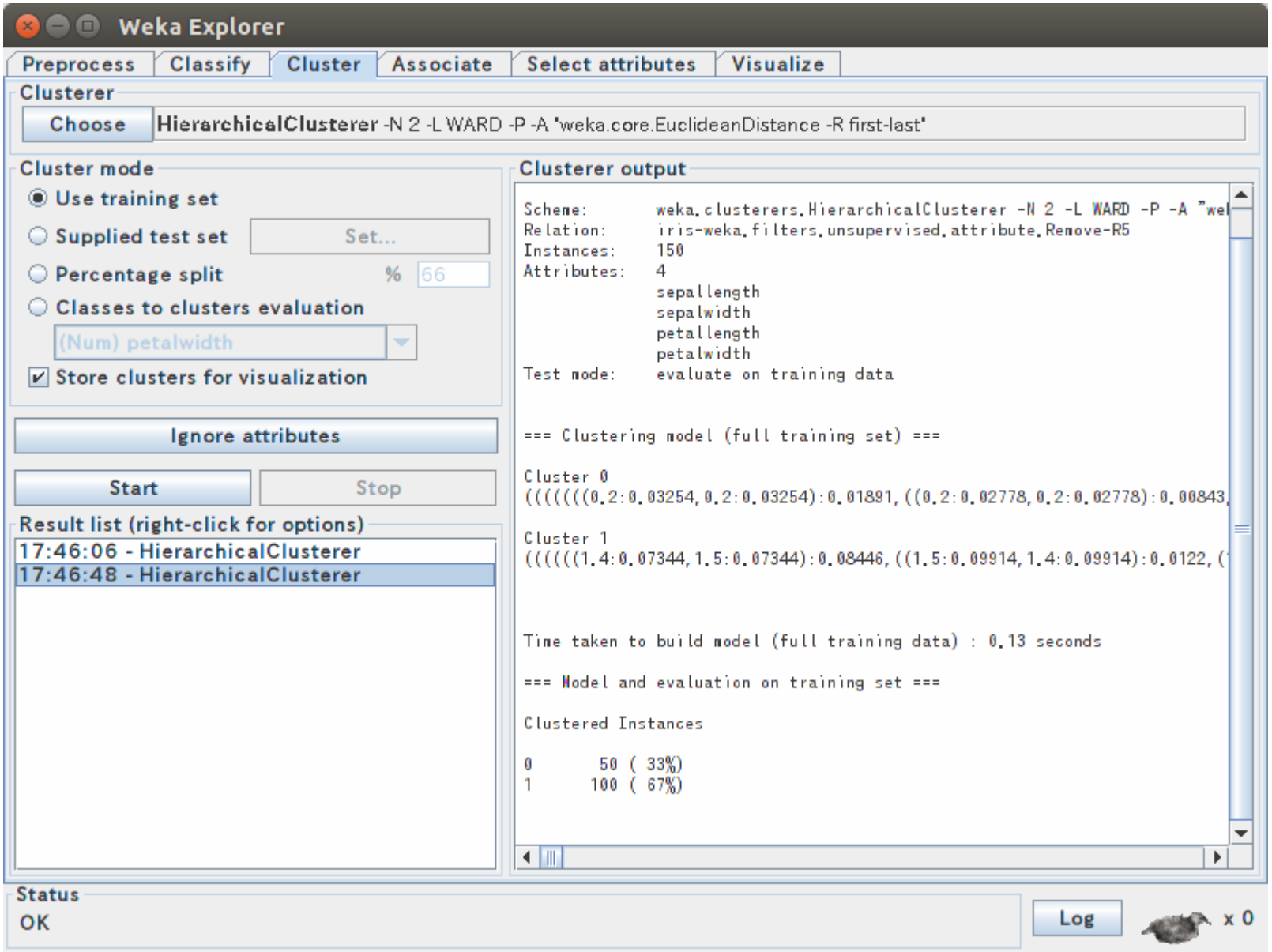
10.4 確率密度推定

- k-means 法の一般化
  - k 個の平均ベクトルを乱数で決める
    ⇒ k 個の正規分布を乱数で決める
  - 平均ベクトルとの距離を基準に、各データをいずれかのクラスタに所属させる
    ⇒各分布が各データを生成する確率を計算し、各クラスタにゆるやかに帰属させる
  - 所属させたデータをもとに平均ベクトルを再計算
    ⇒各データのクラスタへの帰属度に基づき各分布のパラメータ（平均値、共分散行列）を再計算



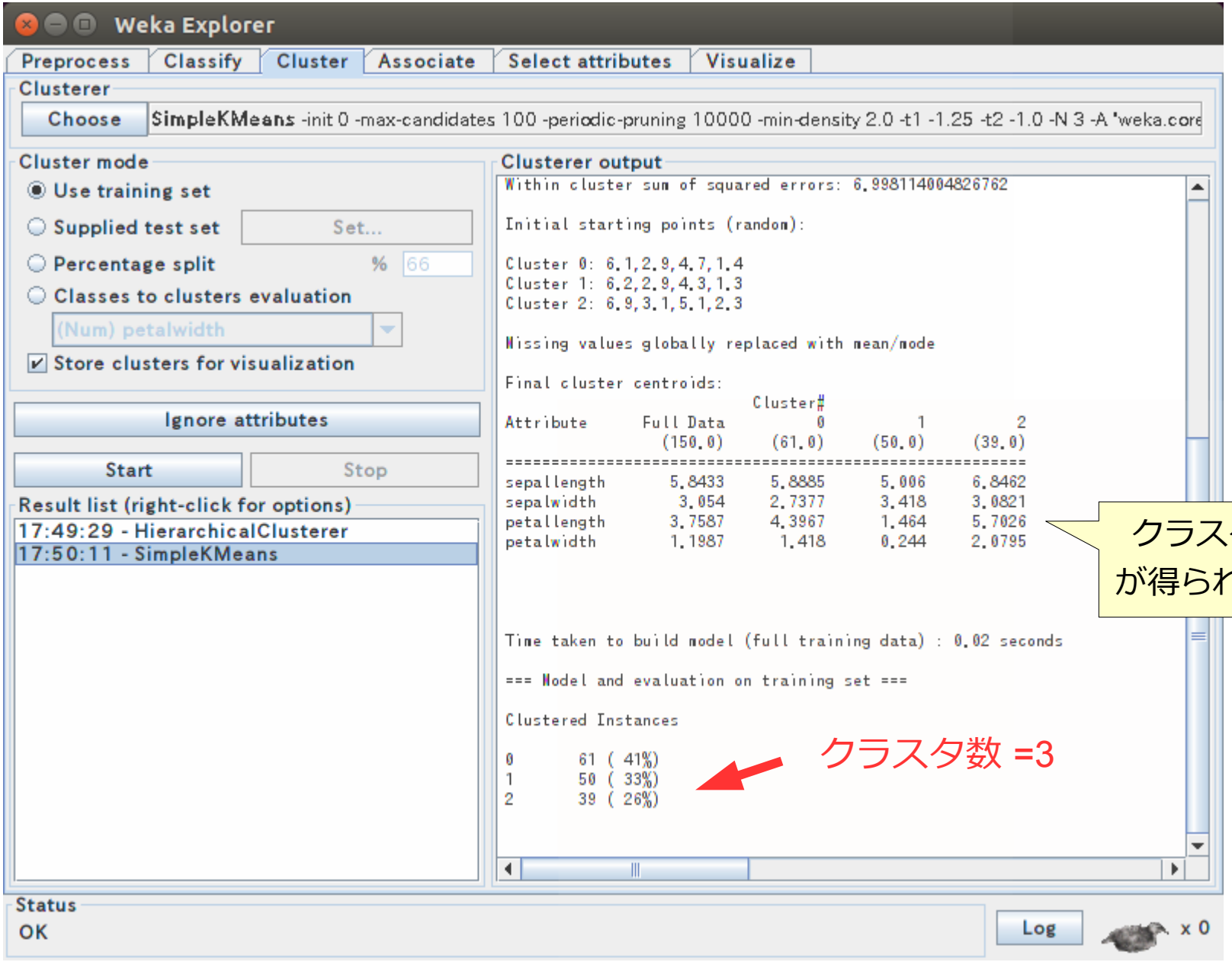
class ラベルを消す

Weka の HierarchicalCluster

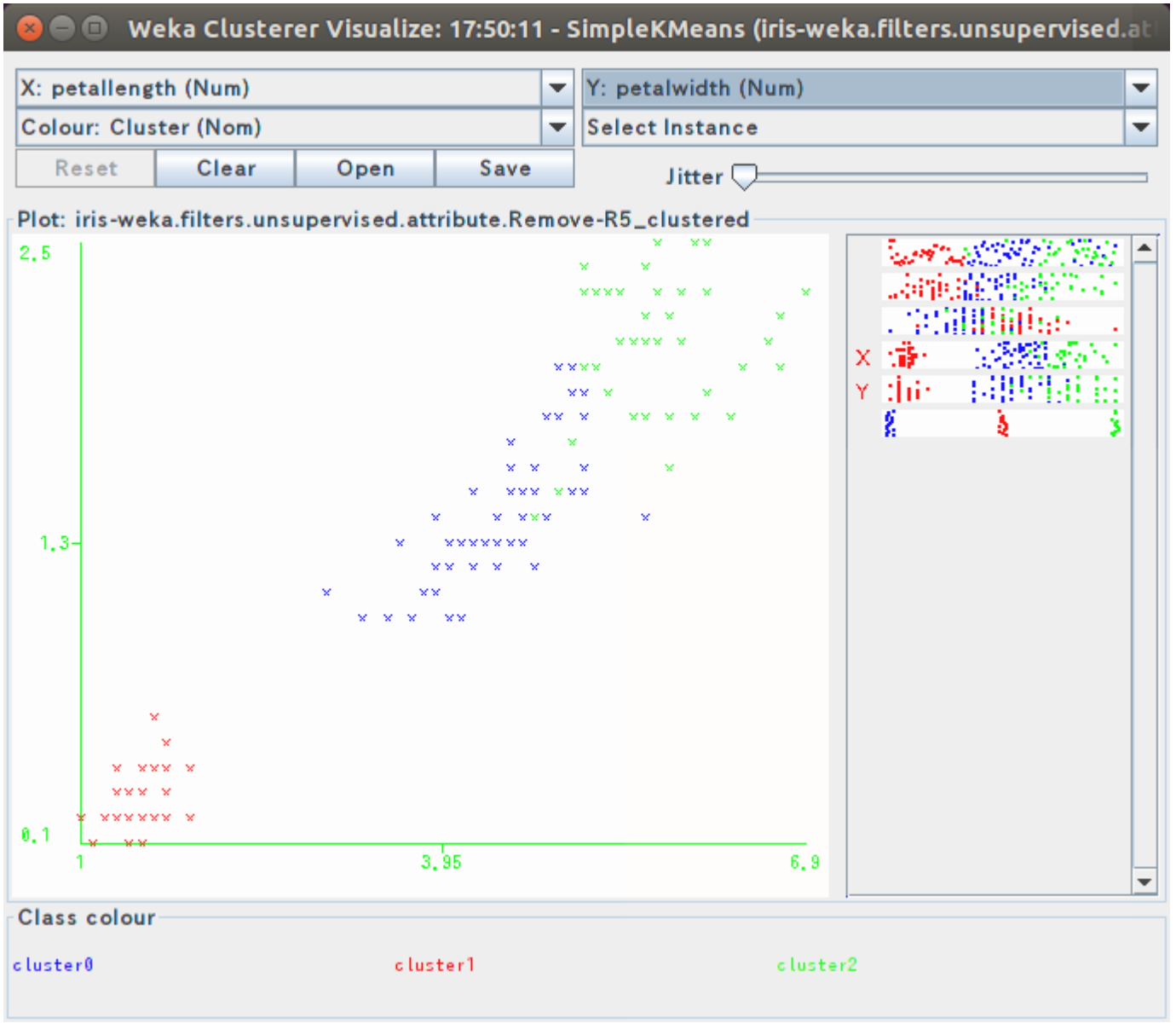


Weka の HierarchicalCluster

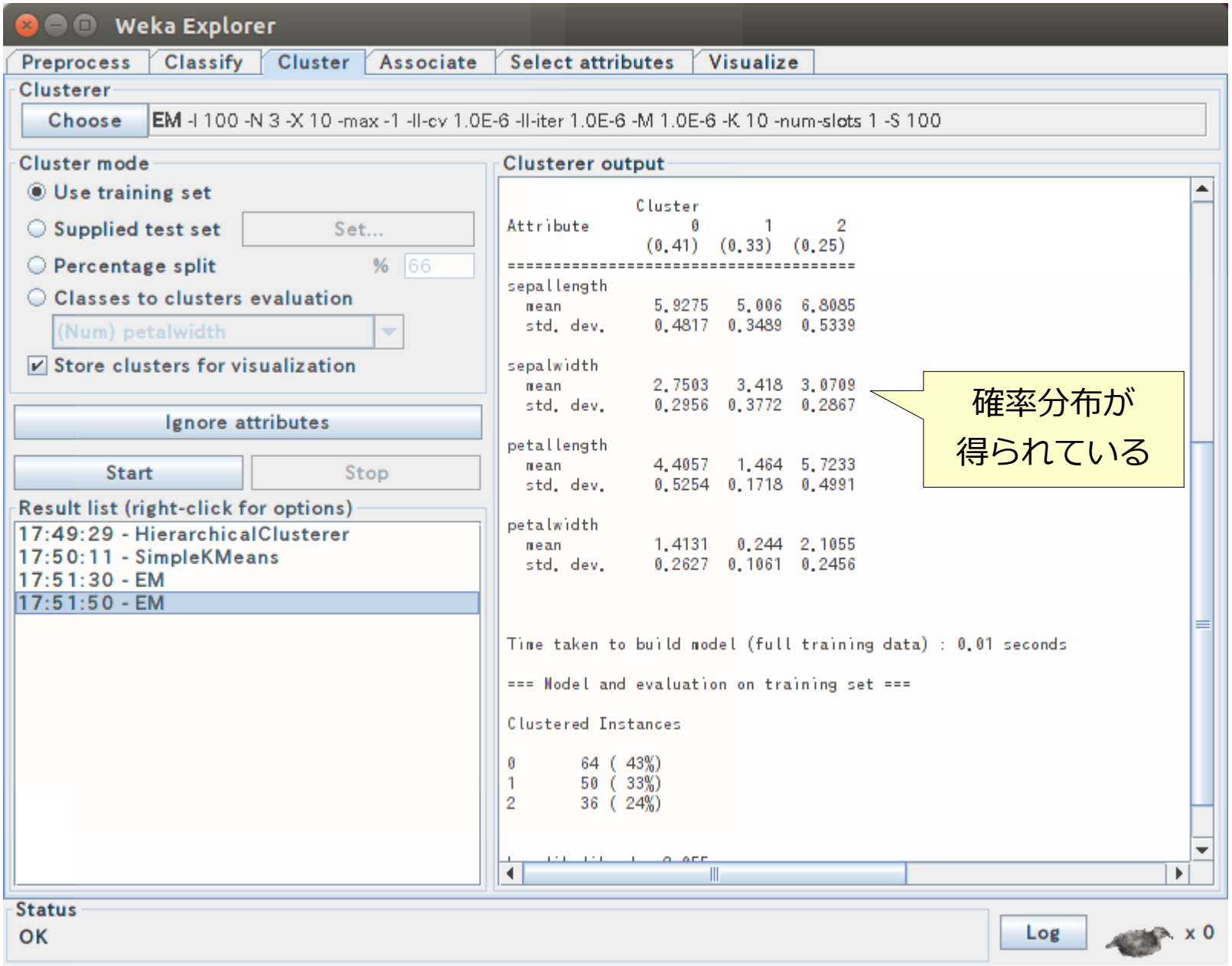
Weka の SimpleKMeans



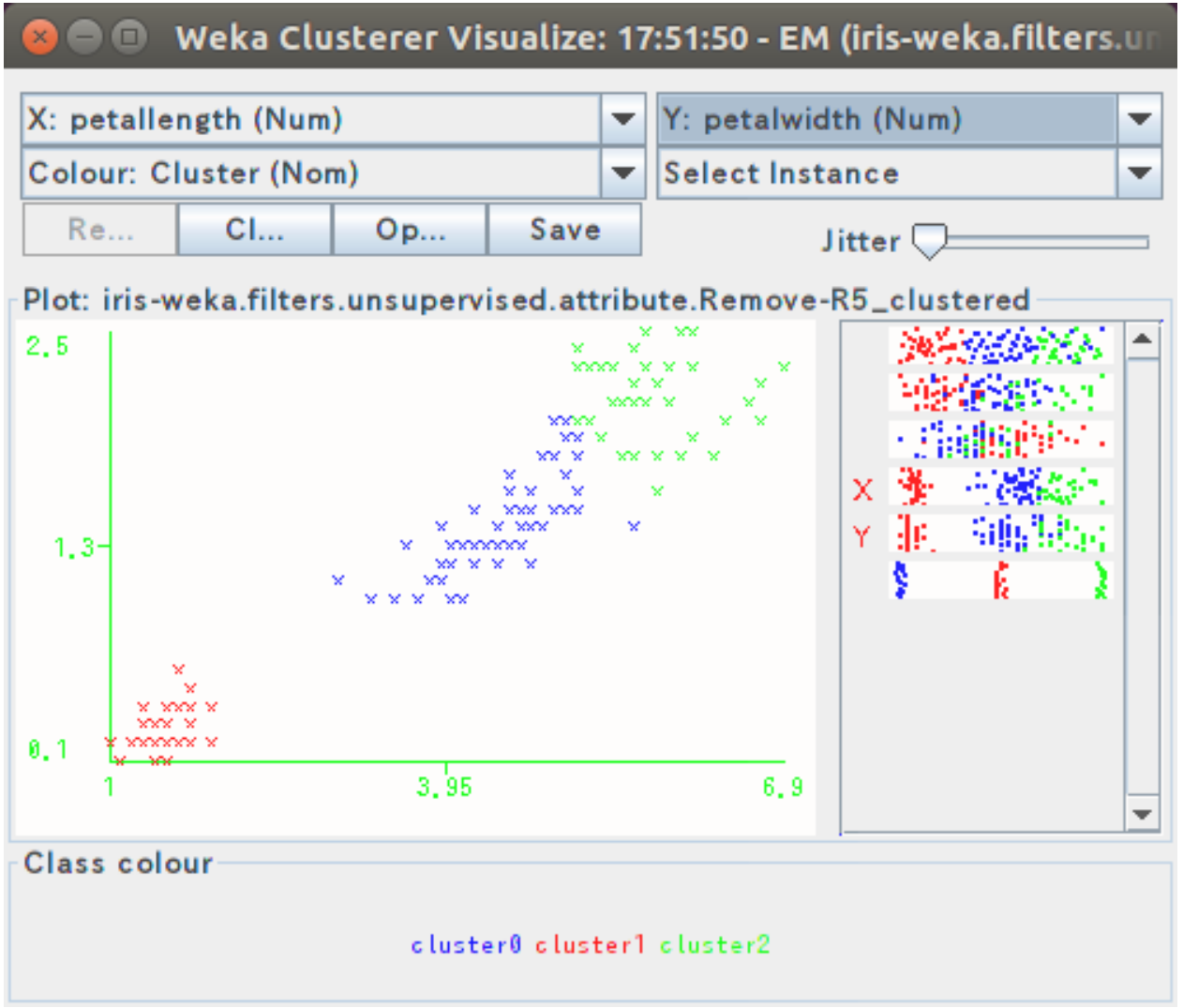
Weka の SimpleKMeans



Weka の EM


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N 3 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) petalwidth
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
17:49:29 - HierarchicalClusterer
17:50:11 - SimpleKMeans
17:51:30 - EM
17:51:50 - EM
Clusterer output
Cluster
Attribute 0 1 2
(0.41) (0.33) (0.25)
sepallength
mean 5.9275 5.006 6.8085
std. dev. 0.4817 0.3489 0.5339
sepalwidth
mean 2.7503 3.418 3.0709
std. dev. 0.2956 0.3772 0.2867
petallength
mean 4.4057 1.464 5.7233
std. dev. 0.5254 0.1718 0.4991
petalwidth
mean 1.4131 0.244 2.1055
std. dev. 0.2627 0.1061 0.2456
Time taken to build model (full training data) : 0.01 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 64 (43%)
1 50 (33%)
2 36 (24%)
確率分布が
得られている
Status
OK
Log
x 0

Weka の EM


Weka Clusterer Visualize: 17:51:50 - EM (iris-weka.filters.un
X: petallength (Num)
Y: petalwidth (Num)
Colour: Cluster (Nom)
Select Instance
Re...
Cl...
Op...
Save
Jitter
Plot: iris-weka.filters.unsupervised.attribute.Remove-R5_clustered
2.5
1.3
0.1
1
3.95
6.9
X
Y
Class colour
cluster0 cluster1 cluster2