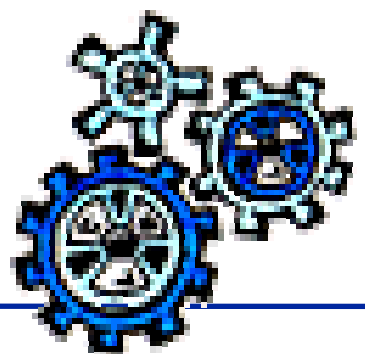
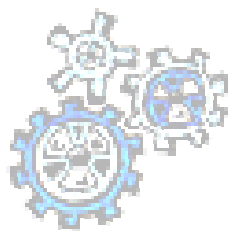
Strumenti per l’Analisi ed il Preprocessing dei dati

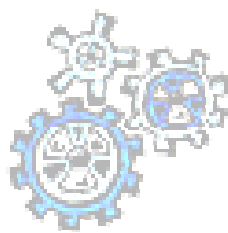
Francesco Folino





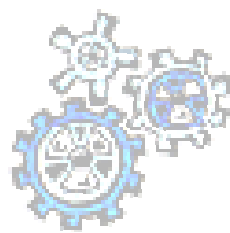
Obiettivo

- **Introdurre gli aspetti essenziali della fase di preparazione dei dati**
- **Acquisire padronanza di un processo tipicamente “artigianale”**
  - **La preparazione dei dati è speculare ad un obiettivo**
  - **La preparazione deve essere adeguata**



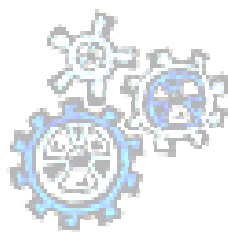
Tools

- **Si adopereranno a tale scopo due strumenti:**
  - **KnowledgeStudio (a scopo dimostrativo)**
  - **Weka (a scopo didattico)**



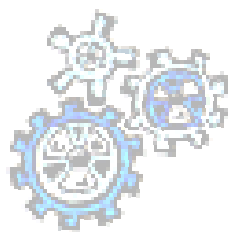
Data Set

- **Il Data Set utilizzato per la dimostrazione è FNBA:**
  - **First National Bank of Anywhere (FNBA)**
  - **Dati relativi a situazioni di credito**
  - **Obiettivo principale: campagna di marketing per attrarre nuovi acquirenti**
    - **La campagna è fatta per gruppi di affinità**
    - **Sulla base delle loro caratteristiche**



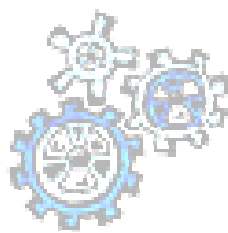
Il Data Preprocessing è un Processo…

- Accesso ai Dati
- Esplorazione dei Dati
  - Sorgenti
  - Quantità
  - Qualità
- Ampliamento e arricchimento dei dati
- Applicazione di tecniche specifiche



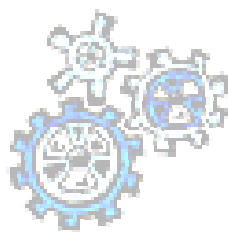
Il Data Preprocessing dipende (ma non sempre) dall'Obiettivo

- Alcune operazioni sono necessarie
  - Studio dei dati
  - Pulizia dei dati
  - Campionamento
- Altre possono essere guidate dagli obiettivi
  - Trasformazioni
  - Selezioni



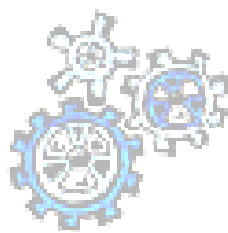
Problemi tipici

- **Troppi dati**
  - dati sbagliati, rumorosi
  - dati non rilevanti
  - dimensione intrattabile
  - mix di dati numerici/simbolici
- **Pochi dati**
  - attributi mancanti
  - valori mancanti
  - dimensione insufficiente



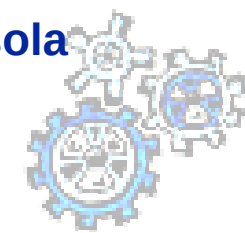
I° passo: Analisi dei Dati

- **Effettuare un'analisi quanto più approfondita sui dati al fine di valutarne al meglio caratteristiche ed eventuali anomalie**
- **A tal fine utilizzeremo:**
  - **Visualizzazione dei dati:**
    - **Distribuzioni**
    - **Diagrammi a barre**
    - **Scatters (o Dot Diagrams per un'analisi della dispersione)**
  - **Misure descrittive dei dati**
    - **Media**
    - **Varianza**
    - **Deviazione Standard**



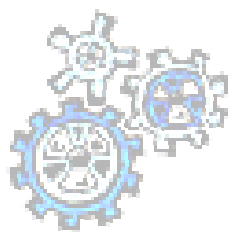
Possibili problemi riscontrabili nei dati

- **Sparsità**
  - Mancanza di valore associato ad una variabile
    - Un attributo è sparso se contiene molti valori nulli
- **Monotonicità**
  - Crescita continua dei valori di una variabile
    - Intervallo [-∞, ∞] (o simili)
  - Non ha senso considerare l'intero intervallo
- **Outliers**
  - Valori singoli o con frequenza estremamente bassa
  - Possono distorcere le informazioni sui dati
- **Dimensionalità**
  - Il numero di valori che una variabile può assumere può essere estremamente alto
    - Tipicamente riguarda valori categorici
- **Anacronismo**
  - Una variabile può essere contingente: abbiamo i valori in una sola porzione dei dati



Analisi dei dati FNBA in Knowledge Studio

- **Obiettivo: analisi preventiva per la riduzione dei dati ad un data set di minore dimensione**
- **I dati devono essere “significativi” e non ridondanti**
- **L’analisi viene fatta sugli attributi e sulle istanze**

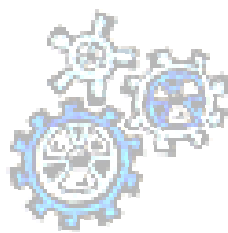


Analisi sugli attributi – Overwiew Report

- **Attributi molto sparsi. Osservando il report, gli attributi:**
  - **EQLIMIT, EQBAL, EQHIGHBAL, EQCURBAL, BCOPEN presentano la maggior parte dei loro valori a NULL.**

- **Presenza di molti attributi outliers:**
  - **CRITERIA, BCOPEN, OWN_HOME, BUYER, EST_INC**

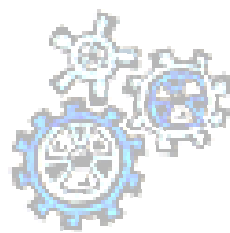
    i cui valori sono singoli o comunque a bassissima frequenza



Analisi sugli attributi – DataSet chart

- **Permette di visualizzare, usando diversi tipi di grafici, le distribuzioni degli attributi.**

- **Consente di avere una conferma immediata sulle ipotesi fatte riguardo agli attributi sotto esame.**

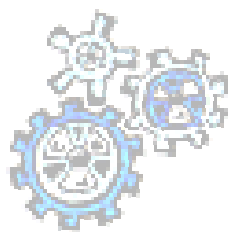


Analisi sugli attributi – Correlations

- **Consente di identificare eventuali correlazioni tra attributi**
- **Ad esempio in FNBA gli attributi MTCURBAL e MTHIGHBAL sono tra loro negativamente correlati (hanno un coefficiente di correlazione pari a 0.98) e come tali non indipendenti. Il coefficiente di correlazione (Pearson) è:**

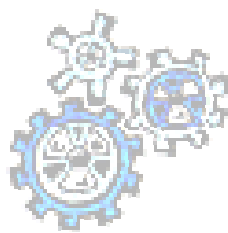
$$r_{xy} = \frac{Cov(x, y)}{s_x s_y}$$

- **Uno tra loro (va scelto opportunamente) è ridondante e come tale può essere eliminato**
- **Passo importante ai fini della riduzione della dimensionalità degli attributi. Inoltre il modello di regressione può essere utile nella sostituzione di valori NULL**



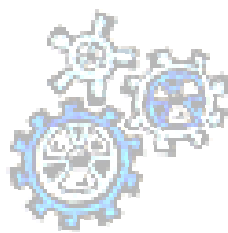
Analisi sugli attributi – Cross Tabs

- **Consente di effettuare diversi tipi di grafici incrociando fra loro gli attributi da analizzare**
- **E' possibile così identificare, qualora esista, una linea di regressione**
- **La curva dovuta all'incrocio fra MTCURBAL e MTHIGHBAL, ad esempio, mostra abbastanza chiaramente il rapporto di dipendenza che esiste tra loro**



Analisi sugli attributi – Segment Viewer

- **Consente di vedere come si distribuiscono gli attributi rispetto ad un fissato valore di un dato attributo.**
- **Può essere utile ad identificare anomalie nei dati: ad esempio attributi che perdono di significato quando altri assumono valori determinati**
- **Può essere utilizzato per identificare istanze poco significative e quindi passibili di eliminazione**



Analisi delle istanze

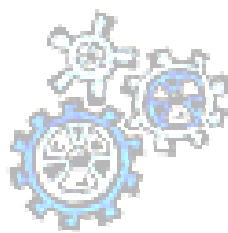
- **Nel caso dell'FNBA Data Set esistono molte istanze poco significative dovute all'altissima presenza di valori NULL**

- **Ad esempio, se segmentiamo tutti gli attributi rispetto al valore NULL assunto dall'attributo RLIMIT, è possibile osservare che molti altri attributi (CRITERIA, ROPEN, RBALNO, RBAL, LST_R_OPEN) contemporaneamente hanno valore NULL.**

- **Si potrebbe così pensare di eliminare tutte le istanze in cui RLIMIT assume valore NULL**

WEKA: il software

- E’ un software scritto in Java e permette di fare analisi di Data Mining
- Usato molto in ambito accademico e di ricerca
- Principali caratteristiche:
  - Contiene un insieme di tools per il data pre-processing e implementazioni di diversi algoritmi di DM
  - Ha un’interfaccia grafica che ne semplifica l’uso
  - Consente di effettuare confronti fra i vari algoritmi messi a disposizione



Weka – Tipo di file

```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal,
    atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
```

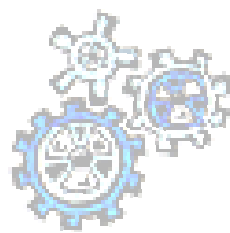
Flat file in ARFF format

Weka – Tipo di file

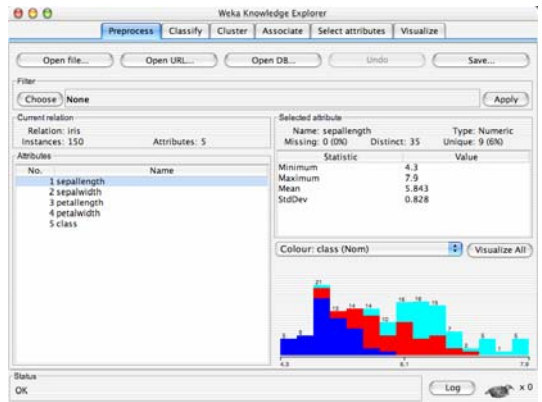
```
@relation heart-disease-simplified

@attribute age numeric
@attribute sex { female, male}
@attribute chest_pain_type { typ_angina, asympt, non_anginal,
    atyp_angina}
@attribute cholesterol numeric
@attribute exercise_induced_angina { no, yes}
@attribute class { present, not_present}

@data
63,male,typ_angina,233,no,not_present
67,male,asympt,286,yes,present
67,male,asympt,229,yes,present
38,female,non_anginal,?,no,not_present
...
```

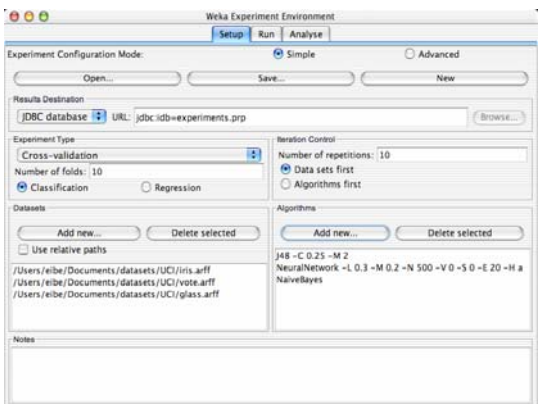
numeric attribute

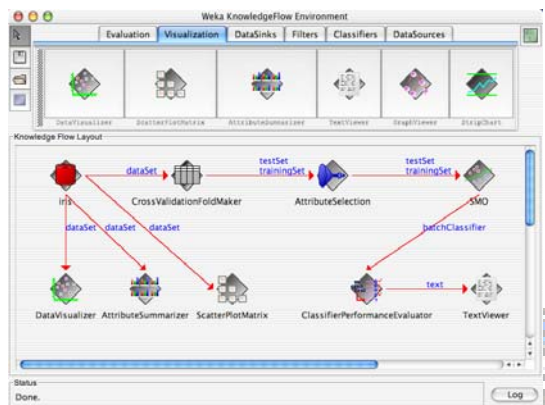
nominal attribute



Weka – l'ambiente

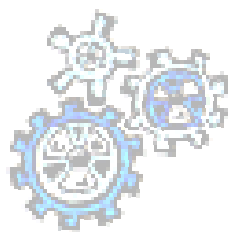


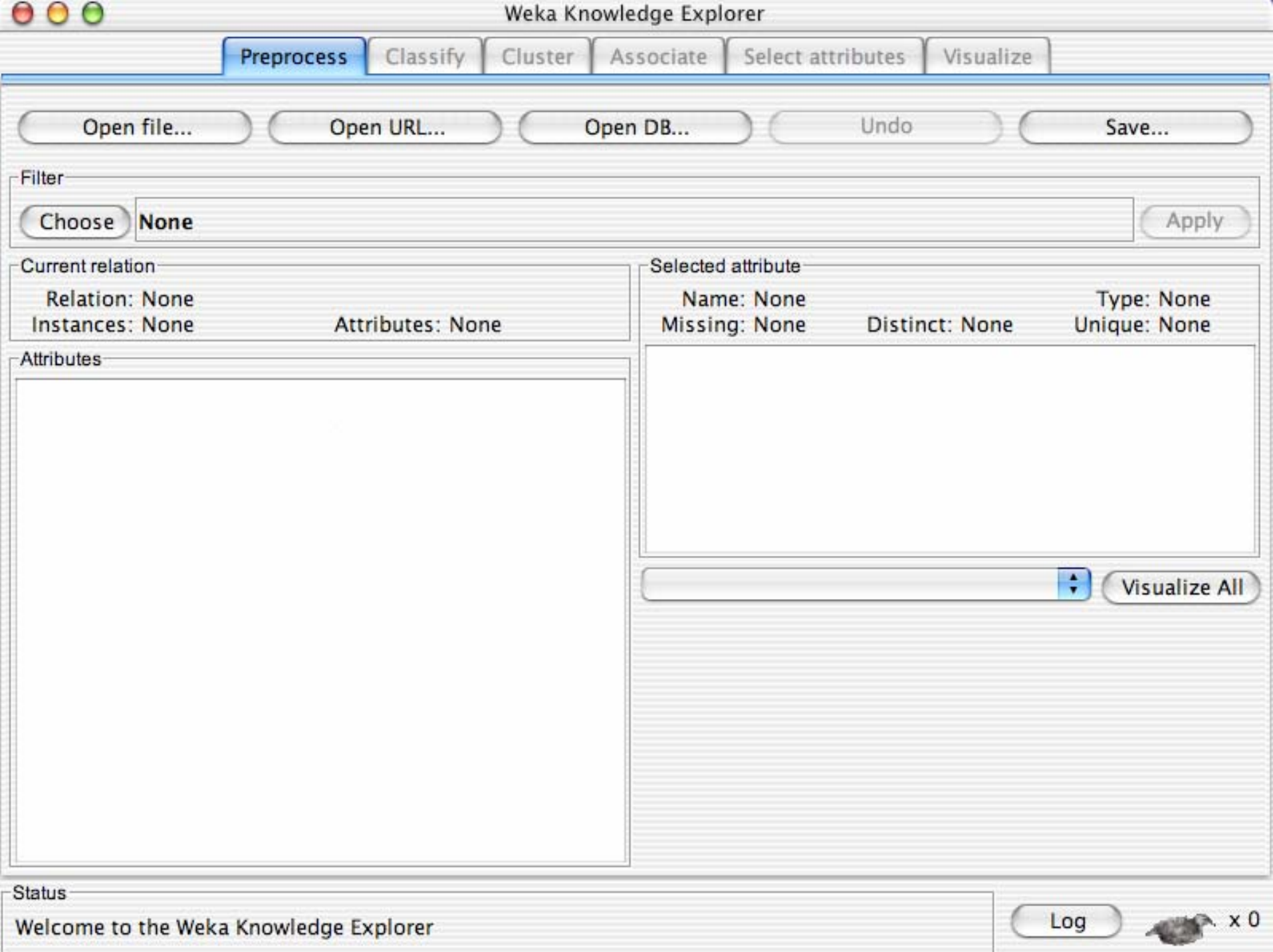


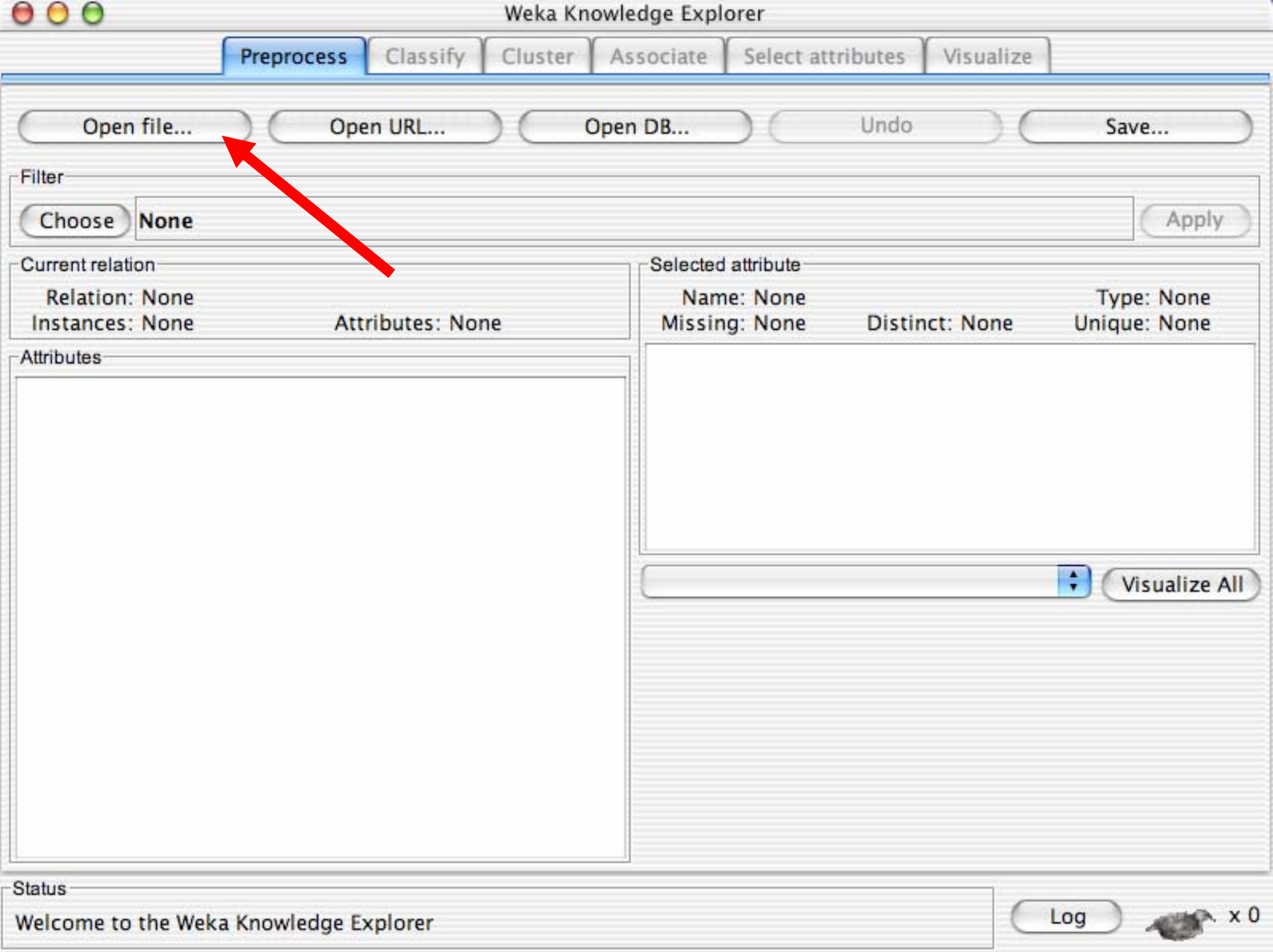


Weka - Explorer

- I dati possono essere importati da files in vari formati: ARFF, CSV, C4.5, binari
- I dati possono essere letti da un URL o da un database (usando JDBC)
- Gli strumenti di pre-processing in WEKA sono chiamati "filtri"
- WEKA contiene filtri per:
  - Discretizzazione, normalizzazione, selezione degli attributi, transformazione di attributi…




Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: None
Instances: None
Attributes: None
Selected attribute
Name: None
Type: None
Missing: None
Distinct: None
Unique: None
Attributes
Visualize All
Status
Welcome to the Weka Knowledge Explorer
Log
x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose **None** | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

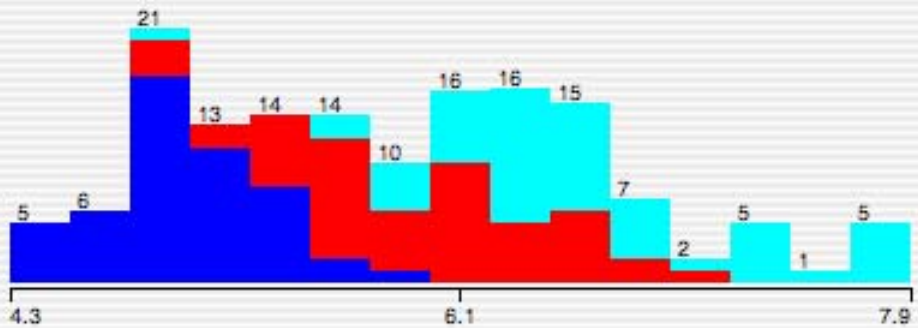
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom) | Visualize All



Status

OK

Log | x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

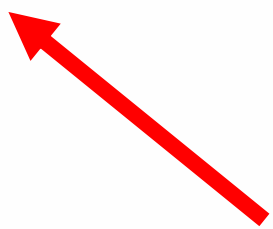
Choose **None** | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

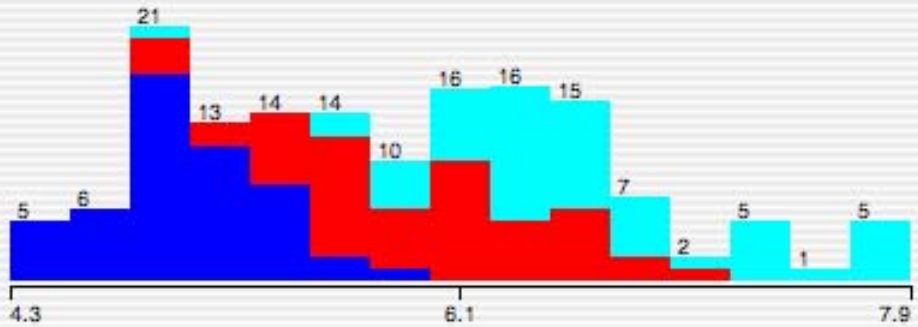


Selected attribute

Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Colour: class (Nom) | Visualize All



Status

OK

Log | x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose **None** | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

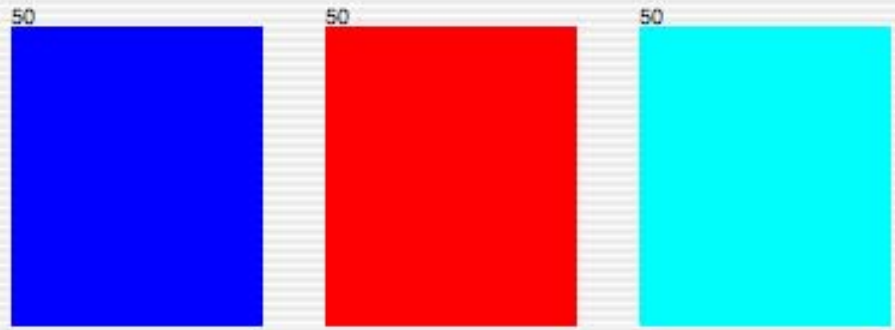
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: class Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Label	Count
Iris-setosa	50
Iris-versicolor	50
Iris-virginica	50

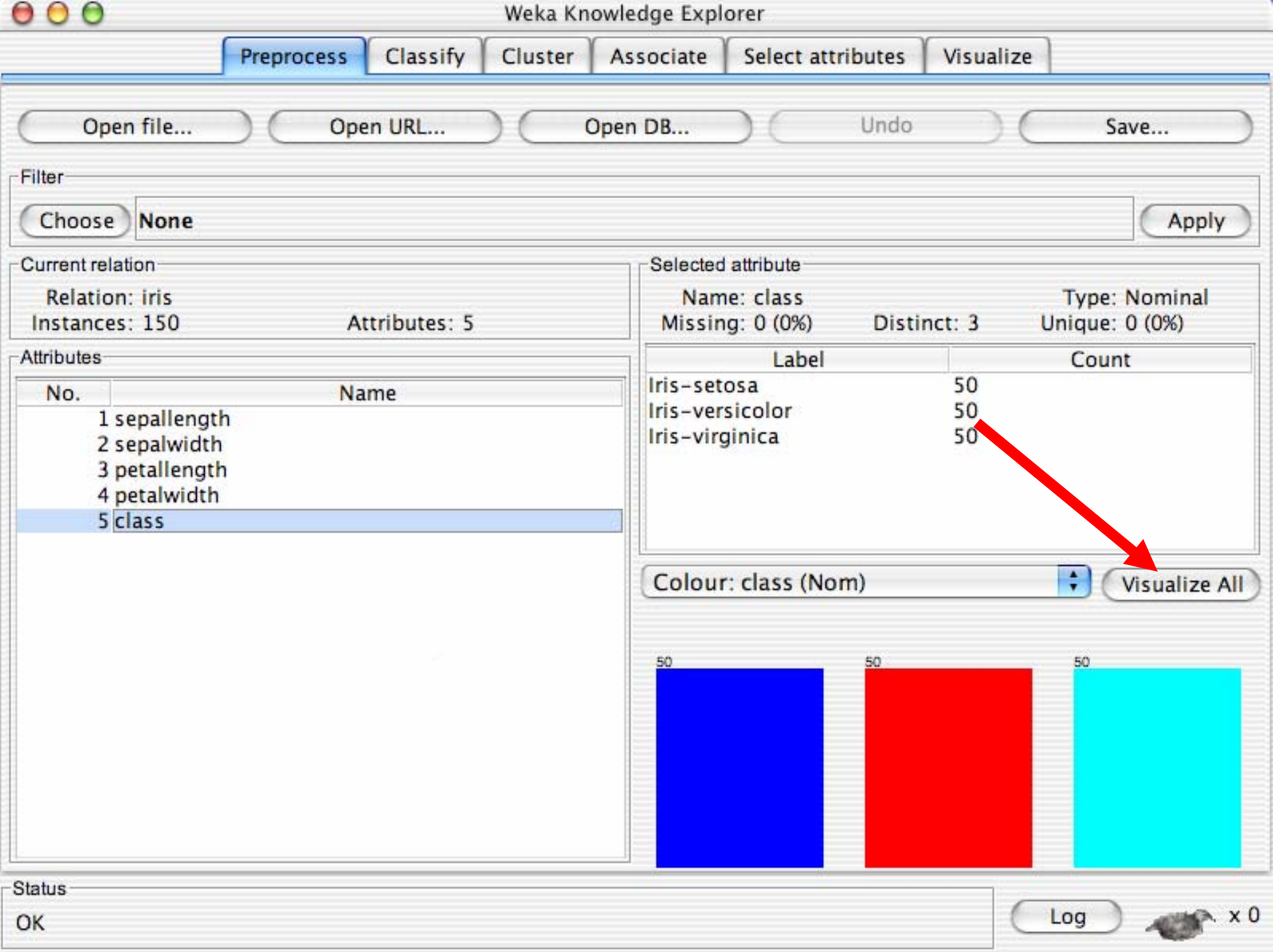
Colour: class (Nom) | Visualize All

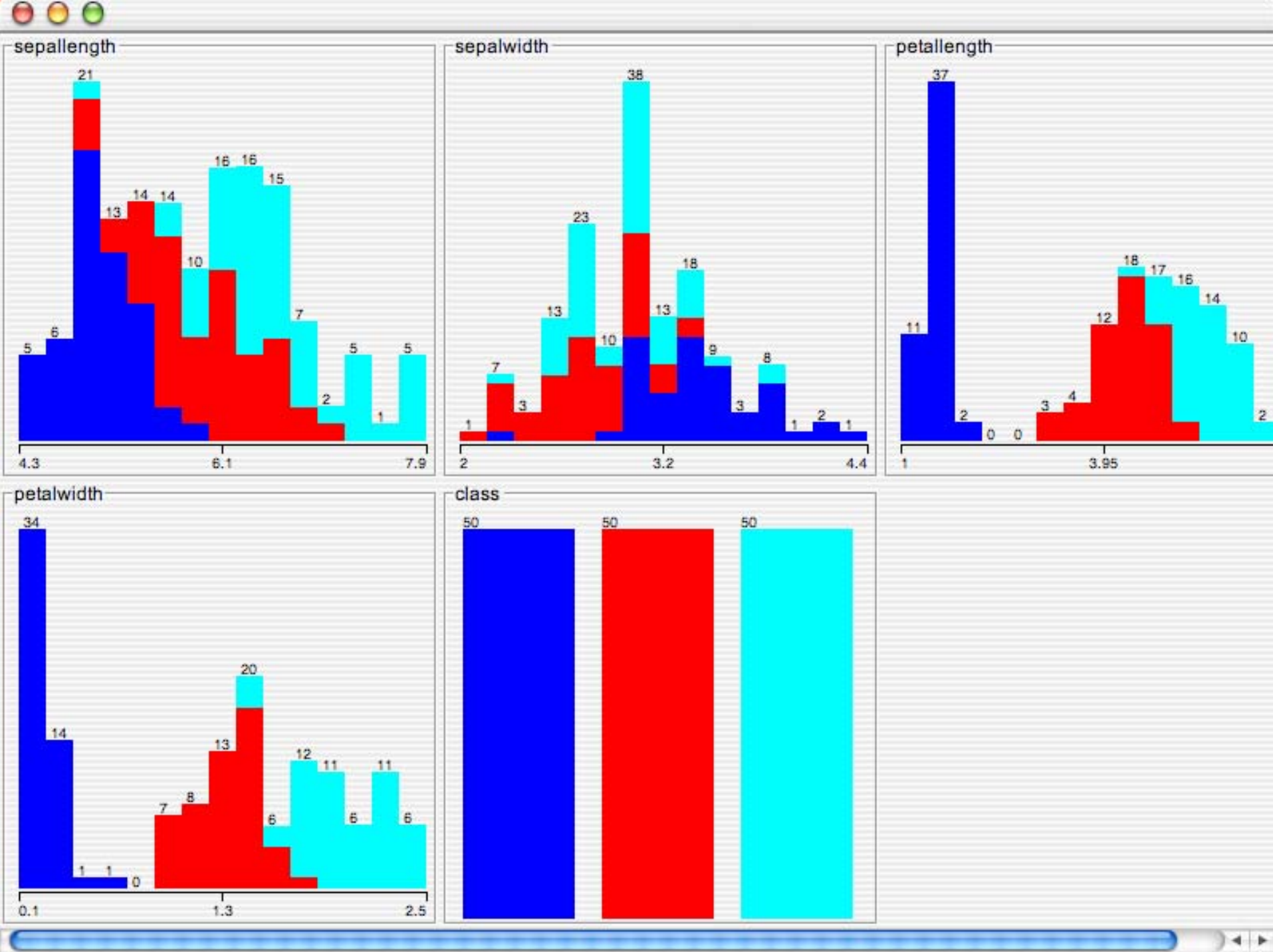


Status

OK

Log | x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Selected attribute
Name: class
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Label
Count
Iris-setosa
50
Iris-versicolor
50
Iris-virginica
50
Colour: class (Nom)
Visualize All
50
50
50
Status
OK
Log
x 0


sepallength
21
13
14
14
10
16
16
15
7
2
5
1
5
5
6
4.3
6.1
7.9
sepalwidth
38
23
13
10
18
13
9
8
3
1
2
1
7
3
1
2
3.2
4.4
petallength
37
11
2
0
0
3
4
12
18
17
16
14
10
2
1
3.95
petalwidth
34
14
1
1
0
20
13
12
11
11
7
8
6
6
6
0.1
1.3
2.5
class
50
50
50

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose **None** | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

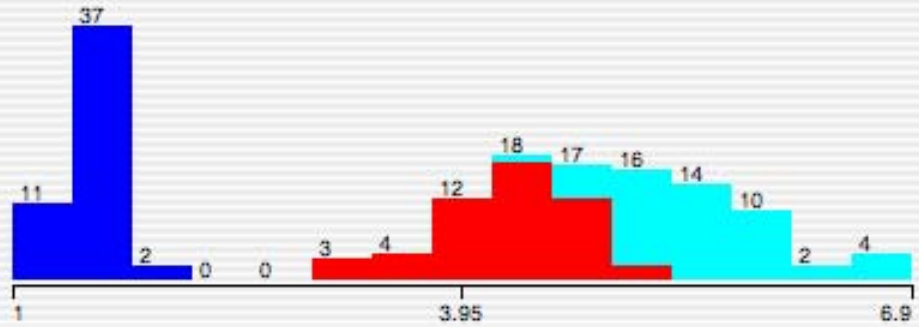
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

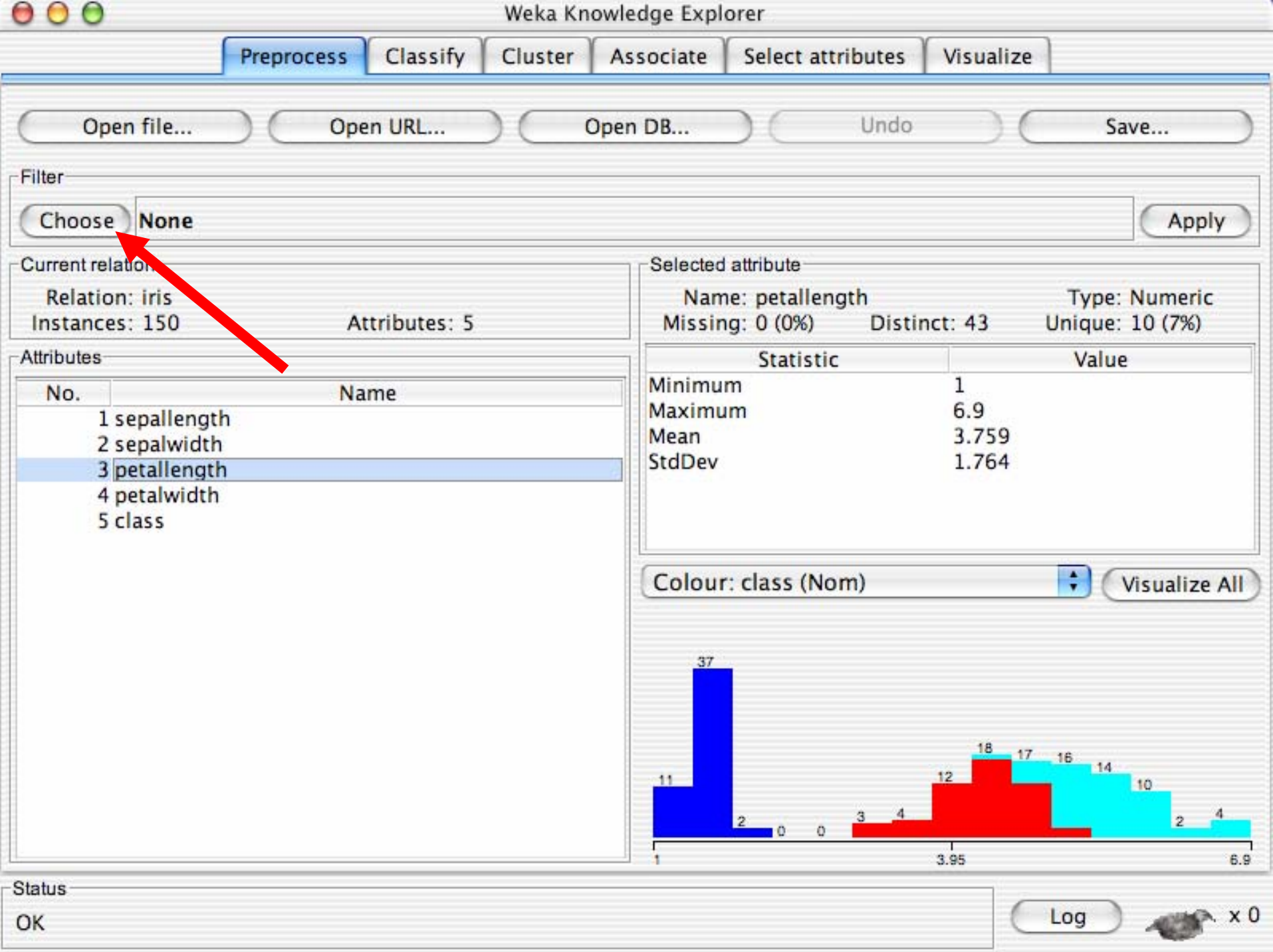
Colour: class (Nom) | Visualize All



Status

OK

Log | x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
None
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

- weka
  - filters
    - unsupervised
      - attribute
      - instance

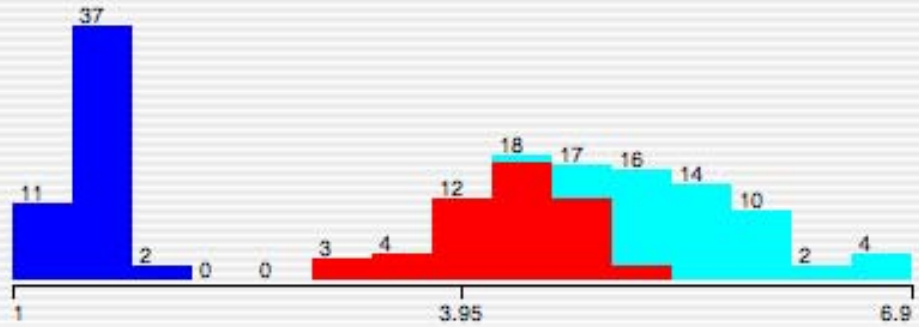
Apply

Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

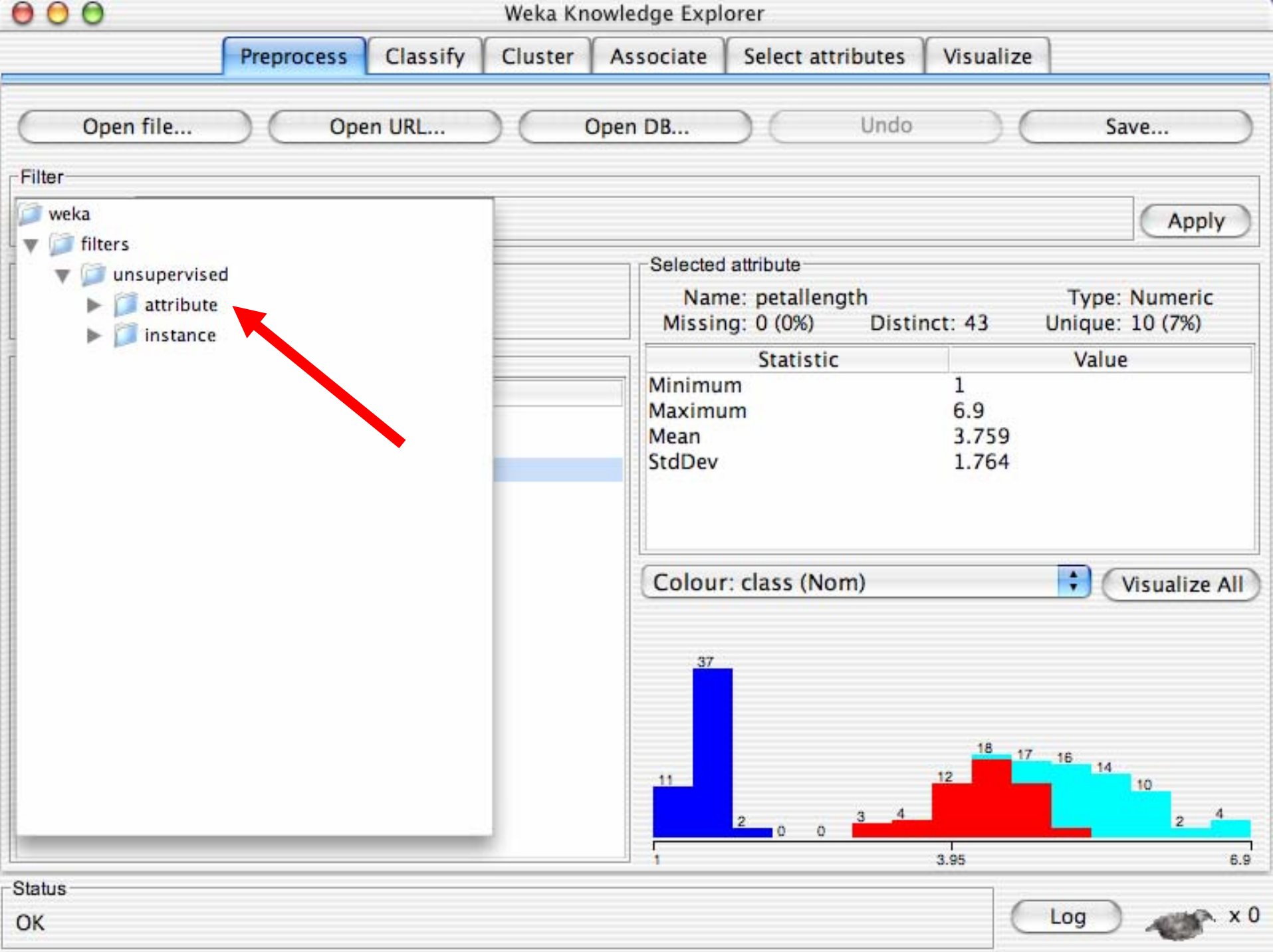
Colour: class (Nom) | Visualize All



Status

OK

Log | x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
weka
filters
unsupervised
attribute
instance
Apply
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

- weka
  - filters
    - unsupervised
      - attribute
        - Add
        - AddCluster
        - AddExpression
        - AddNoise
        - Copy
        - Discretize
        - FirstOrder
        - MakeIndicator
        - MergeTwoValues
        - NominalToBinary
        - Normalize
        - NumericToBinary
        - NumericTransform
        - Obfuscate
        - PKIDiscretize
        - Remove
        - RemoveType

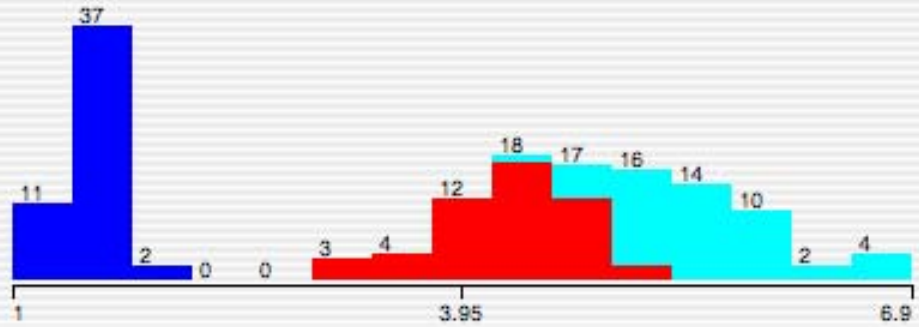
Apply

Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom) | Visualize All



Status

OK

Log | x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose **Discretize** -B 10 -R first-last | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

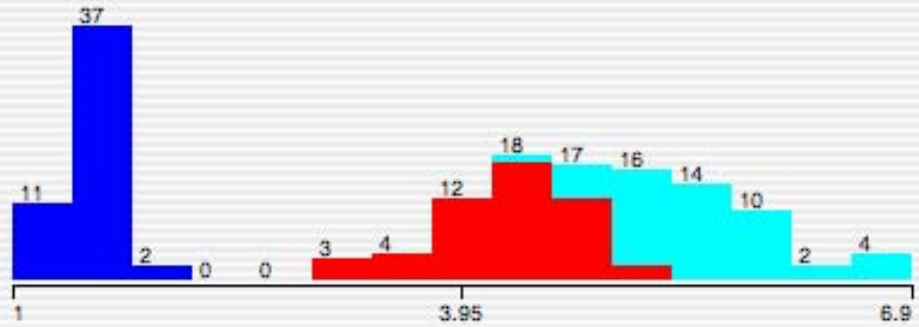
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

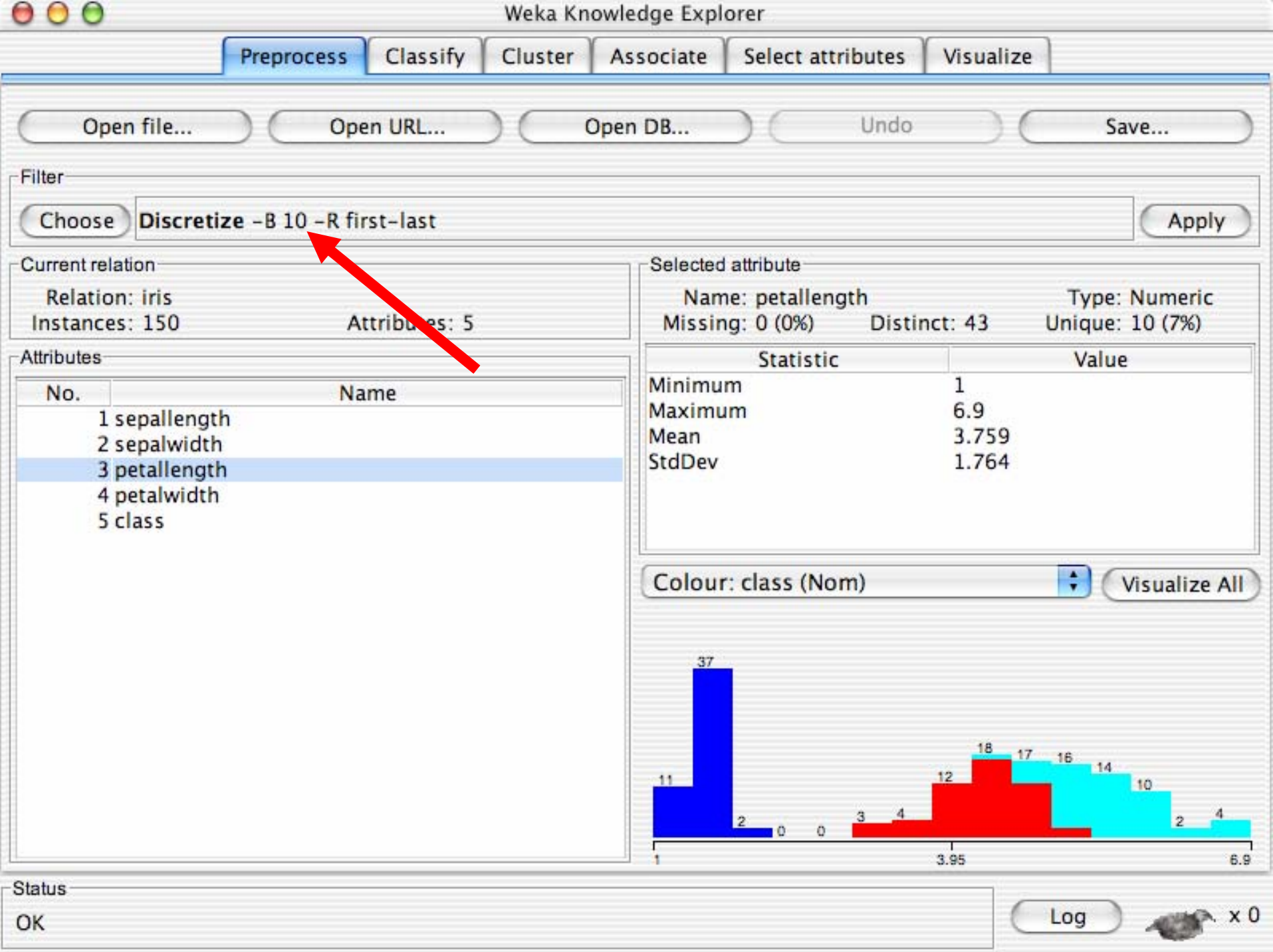
Colour: class (Nom) | Visualize All

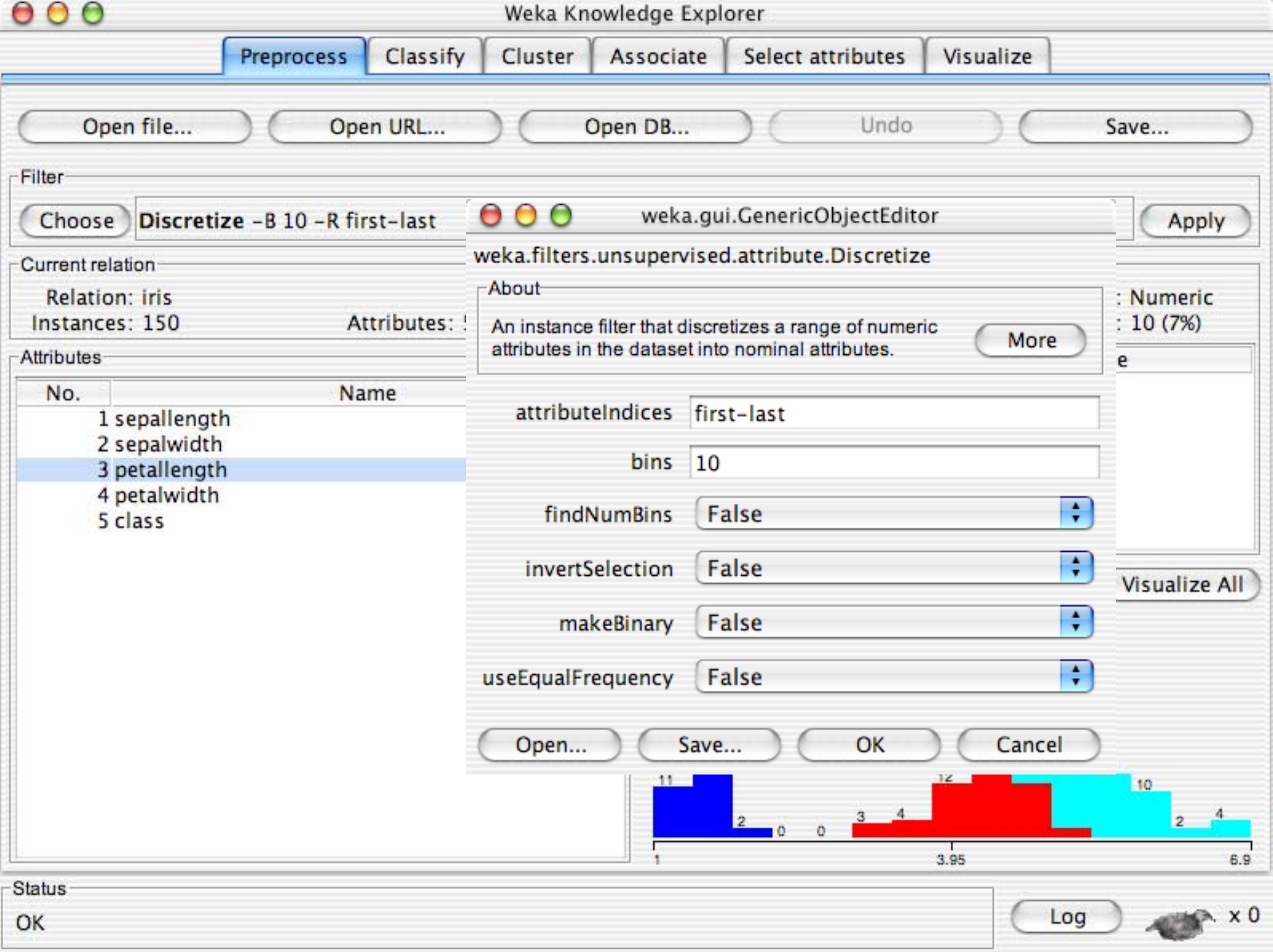


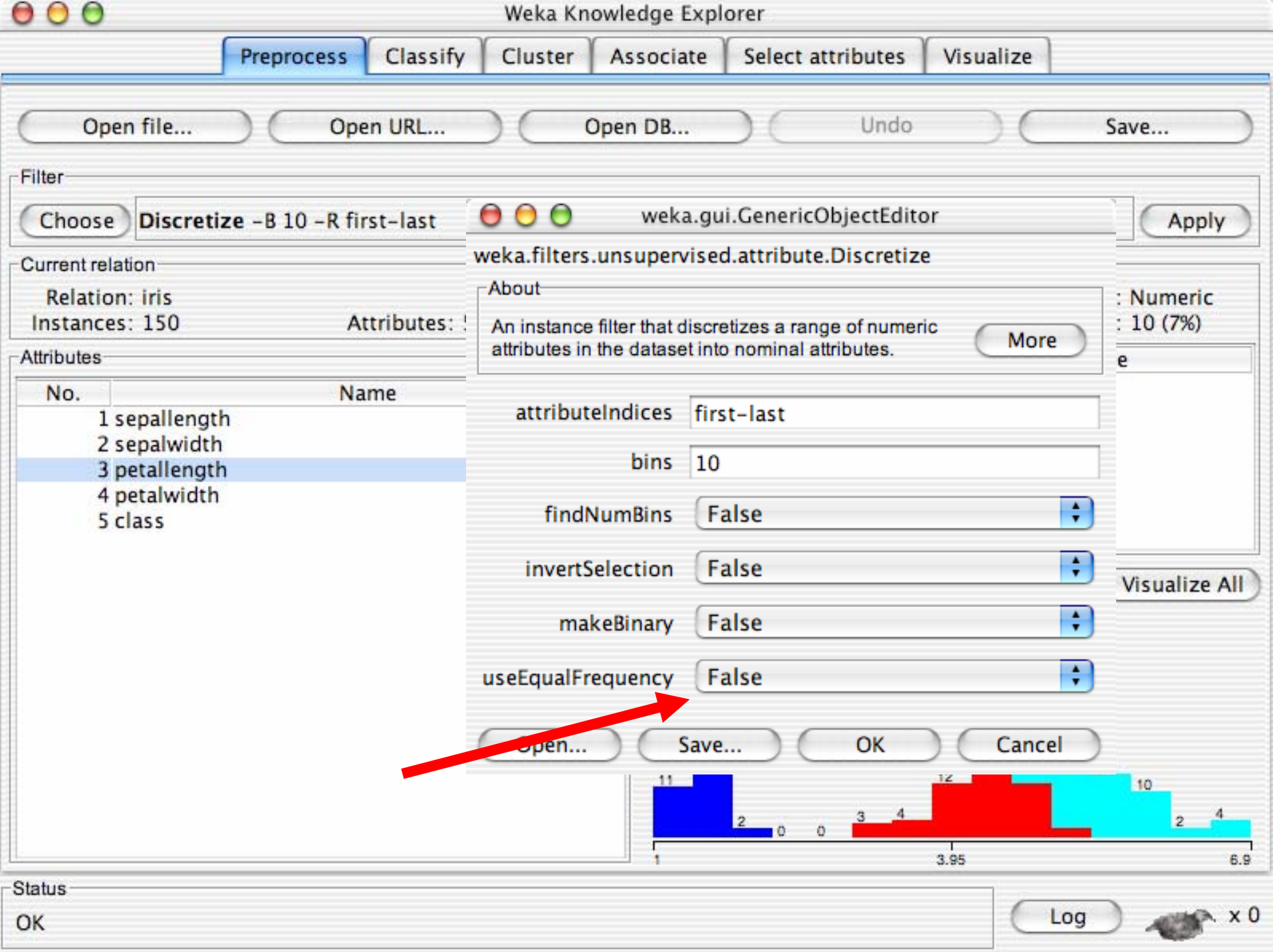
Status

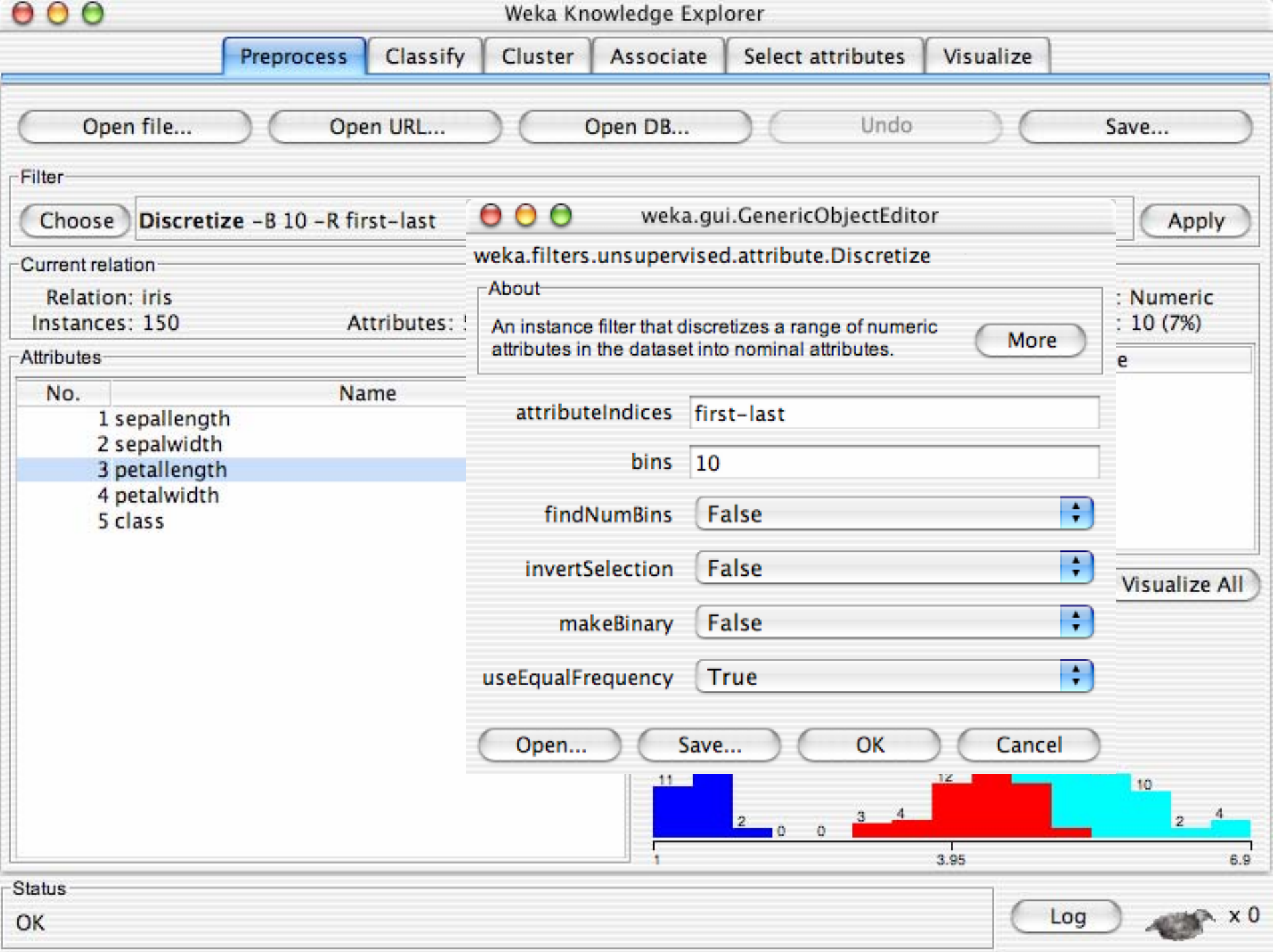
OK

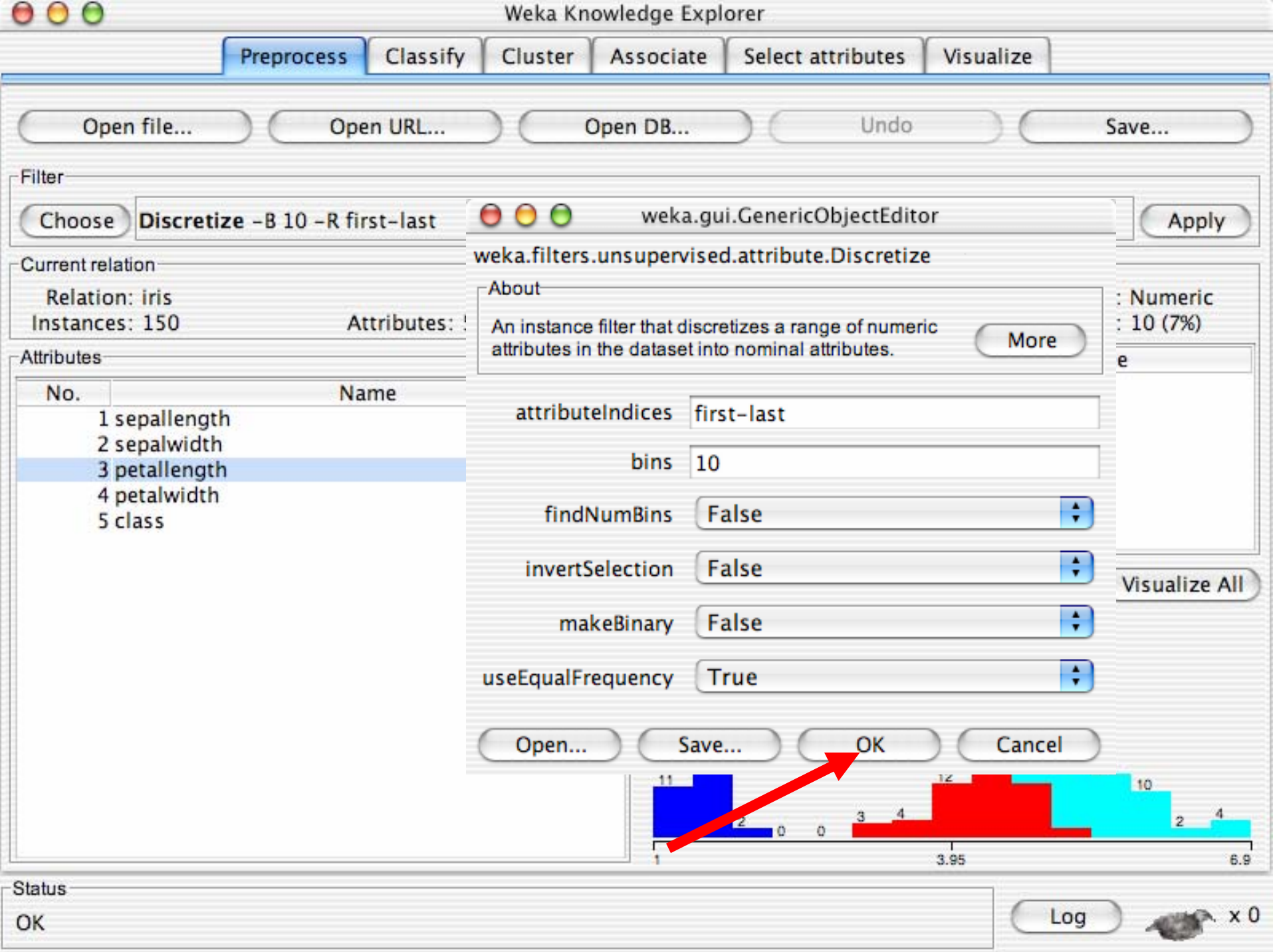
Log | x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes: 5
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Selected attribute
Name: petallength
Type: Numeric
Missing: 0 (0%)
Distinct: 43
Unique: 10 (7%)
Statistic
Value
Minimum
1
Maximum
6.9
Mean
3.759
StdDev
1.764
Colour: class (Nom)
Visualize All
37
11
2
0
0
3
4
12
18
17
16
14
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
False
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
e
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
False
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
e
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Undo
Save...
Filter
Choose
Discretize -B 10 -R first-last
Apply
Current relation
Relation: iris
Instances: 150
Attributes:
Attributes
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
weka.gui.GenericObjectEditor
weka.filters.unsupervised.attribute.Discretize
About
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes.
More
attributeIndices
first-last
bins
10
findNumBins
False
invertSelection
False
makeBinary
False
useEqualFrequency
True
Open...
Save...
OK
Cancel
: Numeric
: 10 (7%)
Visualize All
11
2
0
0
3
4
12
10
2
4
1
3.95
6.9
Status
OK
Log
x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose **Discretize** -F -B 10 -R first-last | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

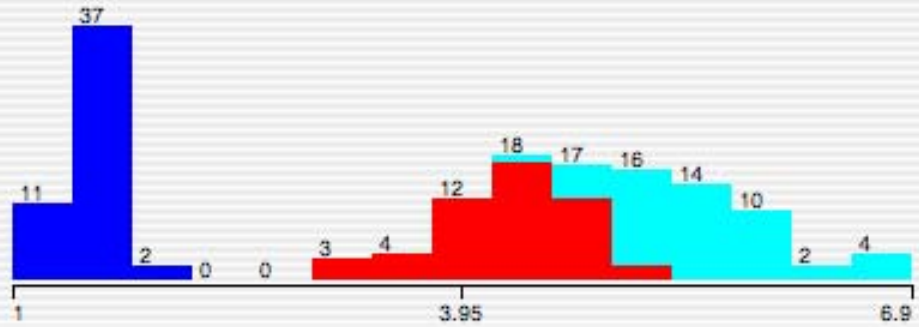
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom) | Visualize All



Status

OK

Log x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose **Discretize** -F -B 10 -R first-last | Apply

Current relation

Relation: iris
Instances: 150 Attributes: 5

Attributes

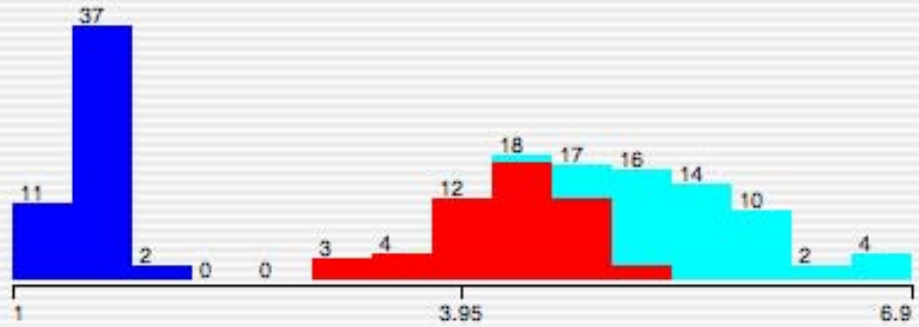
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength Type: Numeric
Missing: 0 (0%) Distinct: 43 Unique: 10 (7%)

Statistic	Value
Minimum	1
Maximum	6.9
Mean	3.759
StdDev	1.764

Colour: class (Nom) | Visualize All



Status

OK

Log x 0

Weka Knowledge Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open file... | Open URL... | Open DB... | Undo | Save...

Filter

Choose **Discretize** -F -B 10 -R first-last | Apply

Current relation

Relation: iris-weka.filters.unsupervised.attribute.Disc...
Instances: 150 Attributes: 5

Attributes

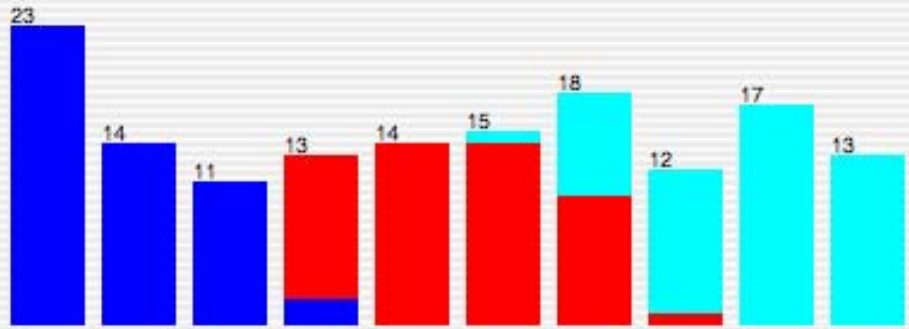
No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

Selected attribute

Name: petallength Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

Label	Count
'(-inf-1.45]'	23
'(1.45-1.55]'	14
'(1.55-1.8]'	11
'(1.8-3.95]'	13
'(3.95-4.35]'	14
'(4.35-4.65]'	15
'(4.65-5.05]'	18

Colour: class (Nom) | Visualize All



Status

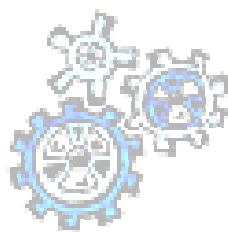
OK

Log | x 0

Trattamento dei dati in Weka

- **L'analisi esplorativa dei dati fatta in Knowledge Studio può essere interamente replicata in Weka usando i tools Preprocess e Visualize nel pannello Explorer.**

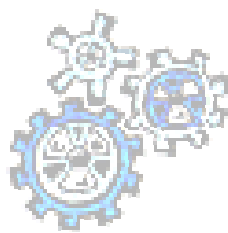
- **Affrontiamo perciò come trattare i problemi evidenziati nell'analisi usando:**
  - **Sostituzione dei valori NULL**
  - **Eliminazione di attributi**
  - **Generazione di nuove variabili**
  - **Normalizzazione**
  - **Discretizzazione**
  - **….**



Trattamento dei dati in Weka

- **Sostituzione dei valori NULL:**
  - **Utilizzando media/mediana/moda**
  - **Predicendo i valori mancanti utilizzando la distribuzione dei valori non nulli**
  - **Segmentando i dati (tramite le distribuzioni di altre variabili) e utilizzando misure statistiche (media/moda/mediana) di ogni segmento**
  - **Costruendo un modello di regressione**

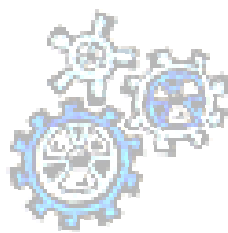
- **In Weka i valori nulli vengono sostituiti con le medie e le mode calcolate sui dati di training (*ReplaceMissingValues* filter)**



Trattamento dei dati in Weka

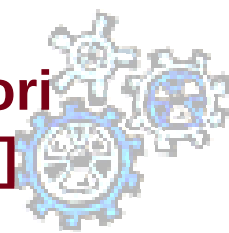
- **Eliminazione degli attributi necessaria allorchè:**
  - **Ci siano moltissimi valori NULL**
  - **Ci siano pochissimi valori distinti**
  - **Esista una correlazione con altri attributi**

- **In Weka tale funzione è svolta dal filtro *Remove* che cancella uno specificato set di attributi dal Data Set di partenza**



Trattamento dei dati in Weka

- La Normalizzazione è utile qualora ci siano:
  - Errori nei dati
  - Dati incompleti
  - Forte asimmetria nei dati
    - diversi raggruppamenti esprimono comportamenti differenti
  - Molti picchi
    - residui larghi e sistematici nella definizione di un modello
- La modifica della forma dei dati può alleviare questi problemi
- In Weka tale funzione è svolta dal filtro *Normalize.* I valori risultanti da questo passo sono valori compresi tra [0,1]



Normalizzazioni

- **min-max normalization**

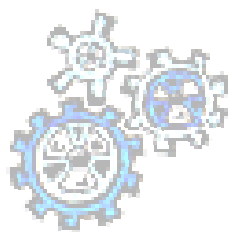
$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- **z-score normalization**

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

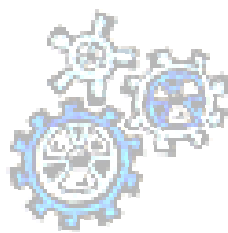
- **normalization tramite decimal scaling**

$$v' = \frac{v}{10^j}$$ dove j è il più piccolo intero tale che Max($|v'|$)<1



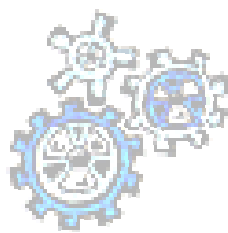
Trattamento dei dati in Weka

- **Discretizzare perché?**
  - **I dati originali possono avere valori continui estremamente sparsi**
  - **I dati discretizzati possono essere più semplici da interpretare**
  - **Le distribuzioni dei dati discretizzate possono avere una forma “Normale”**
- **Esistono due modalità di discretizzazione:**
  - **Supervisionata**
  - **Non supervisionata**



Trattamento dei dati in Weka

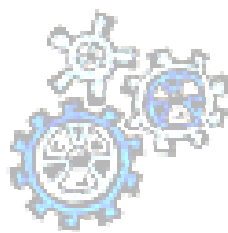
- **Discretizzazione non supervisionata:**
  - **Discretizza senza un preciso criterio (GAIN, ENTROPIA…)**
  - **Il numero di classi è noto a priori**
  - **Natural binning**
    - **intervalli di identica ampiezza**
  - **Equal Frequency binning**
    - **intervalli di identica frequenza**
  - **Statistical binning**
    - **Utilizzando informazioni statistiche**
      - media e varianza
      - Quartili



Trattamento dei dati in Weka

- **Discretizzazione supervisionata:**
  - **La discretizzazione ha un obiettivo quantificabile (entropia, guadagno)**
  - **Il numero di classi non è noto a priori**

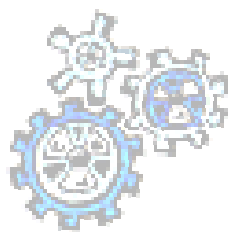
- **I dati discretizzati possono essere ancora estremamente sparsi. In tal caso:**
  - **Eliminazione della variabile in oggetto**



Trattamento dei dati in Weka

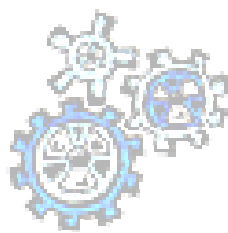
- **Discretizzazione supervisionata in Weka:**
  - **Un filtro che discretizza un range di attributi numerici in attributi nominali. La discretizzazione utilizzata è quella di Fayyad & Irani**

- **Discretizzazione non supervisionata in Weka:**
  - **La discretizzazione è sia natural che frequancy binning.**



Trattamento dei dati in Weka

- La dimensione del Data Set è un problema importante nel Data Mining.
- In Weka è possibile ottenere un sottoinsieme del Data Set originale utilizzando il filtro *Resample*.
- Il Data Set ridotto è ottenuto attraverso una scelta random sulle sue istanze.
- E' importante valutare che il Data Set così generato non snaturi il Data Set di origine.
- A tal fine è possibile osservare la "forma" delle distribuzioni degli attributi del sottoinsieme rispetto a quelle del Data Set originale



Espandibilità di Weka

- **Weka è un ambiente scritto in Java e completamente estendibile.**
- **E' possibile scriversi i propri metodi di filtering, clustering o classificazione e poterli utilizzare nell'ambiente.**
- **Per far ciò occorre conoscere la struttura del codice di Weka.**
- **A fine didattico è stato scritto un filtro *NullHandle* che va a cancellare le istanze nelle quali i valori NULL superano una percentuale fissata**