Classification of Instant Messaging Communications for Forensics Analysis

Angela Orebaugh

Jeremy Allnutt

The International Journal of
FORENSIC COMPUTER SCIENCE (2009)

Presented By:

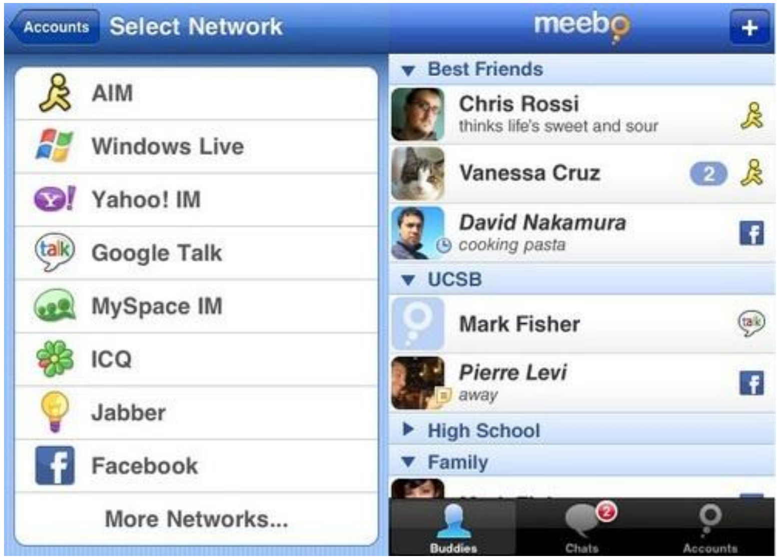
Azzat Ahmed

Outline

- Introduction.
- Objectives.
- Instant Messaging Facts.
- IM Architecture.
- Stylometric Features.
- Experiments and Results.
- Summary & Future work.
- Weka (Parts used in this paper)

Introduction

- Instant Messaging (IM) allows the user to communicate in real time with other users who have the same IM application.
- Falls into a groupware category, i.e. people work together while located remotely.



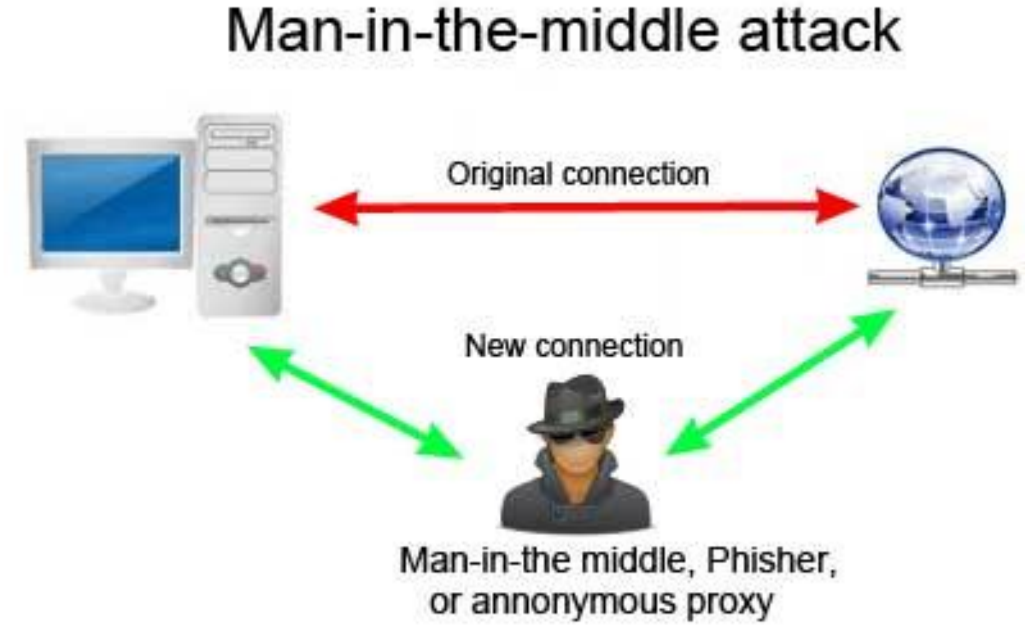
AIM
Windows Live
Yahoo! IM
Google Talk
MySpace IM
ICQ
Jabber
Facebook
More Networks...

Best Friends
Chris Rossi
thinks life's sweet and sour
Vanessa Cruz 2
David Nakamura
cooking pasta
UCSB
Mark Fisher
Pierre Levi
away
High School
Family

Introduction (Cont.)

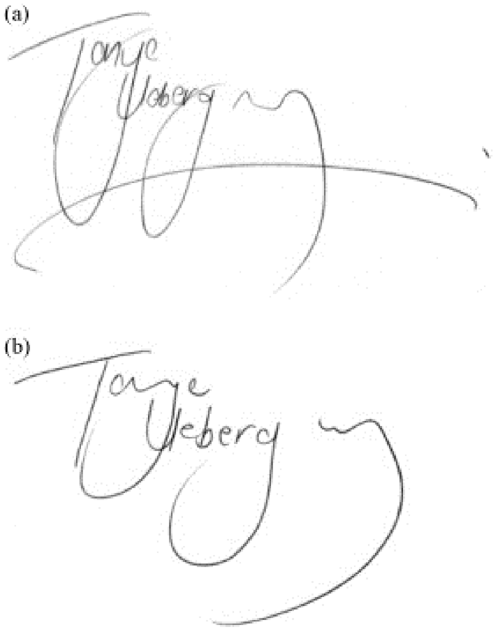
- IM is used widely between people and enterprises
- IM could me misused by attackers.
- Attackers may steal the identity of IM author.

(physically or by hijacking a connection).



Introduction (Cont.)

- Humans have unique patterns of behavior.
- This behavior identifying person.
- IM messages contain unique and constant behavior like biometric data.

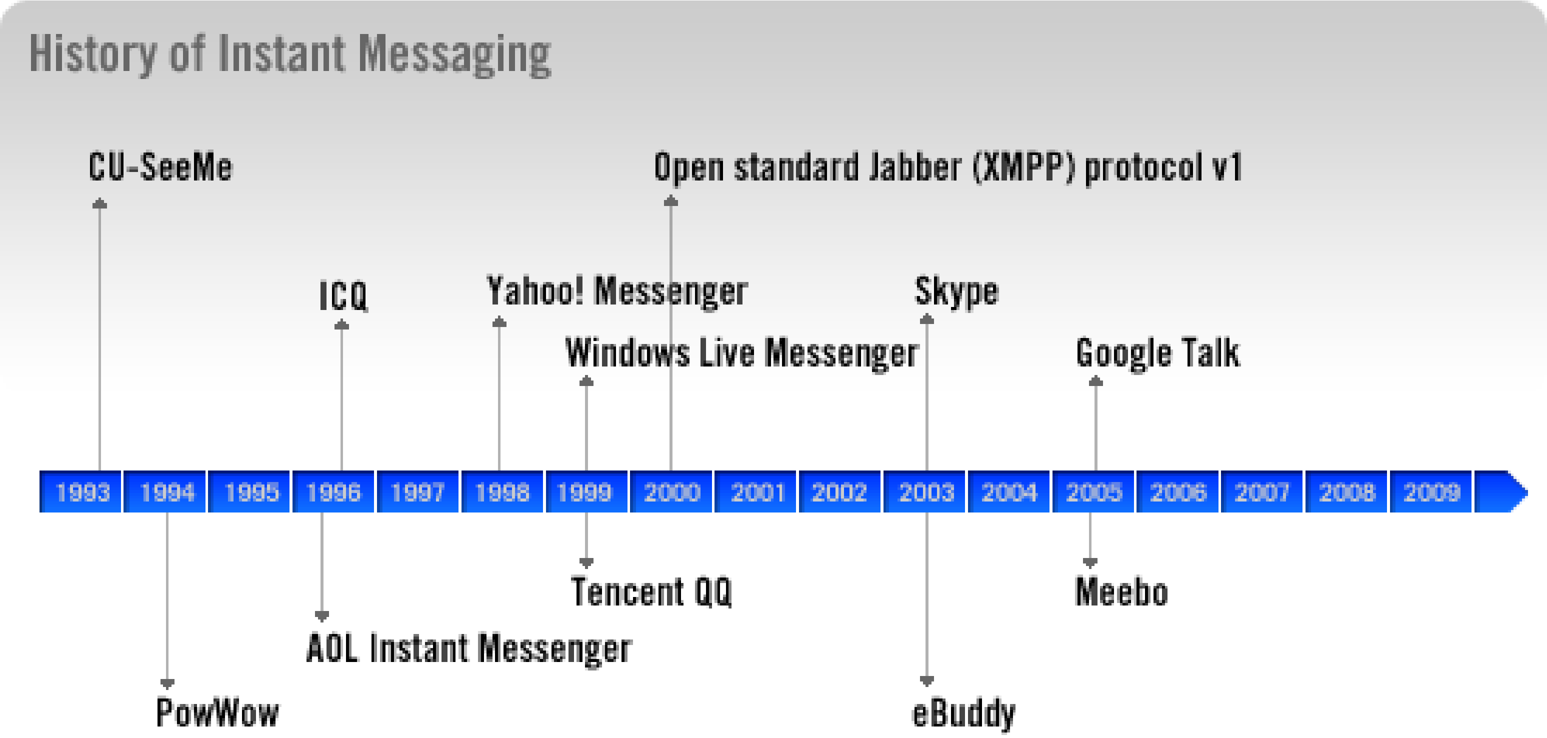


Objectives

- Analysis of Instant Messaging (IM) in terms of digital forensics and intrusion detection.
- Explores IM author classification methods based on author behavior.
- Identification/validation of IM authors for forensics analysis using data mining classification.

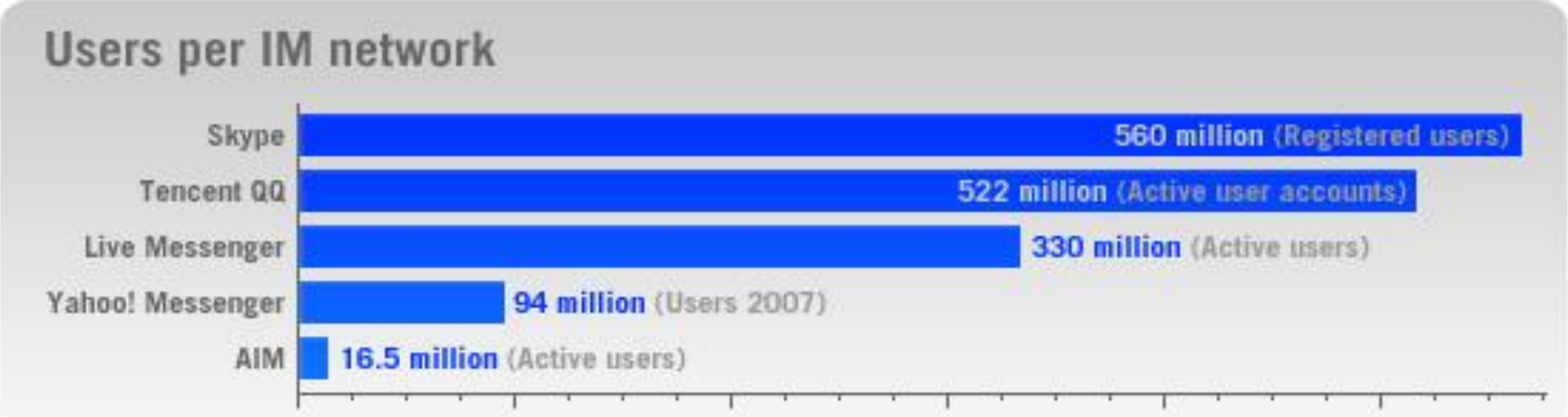
Instant Messaging Facts

Timeline

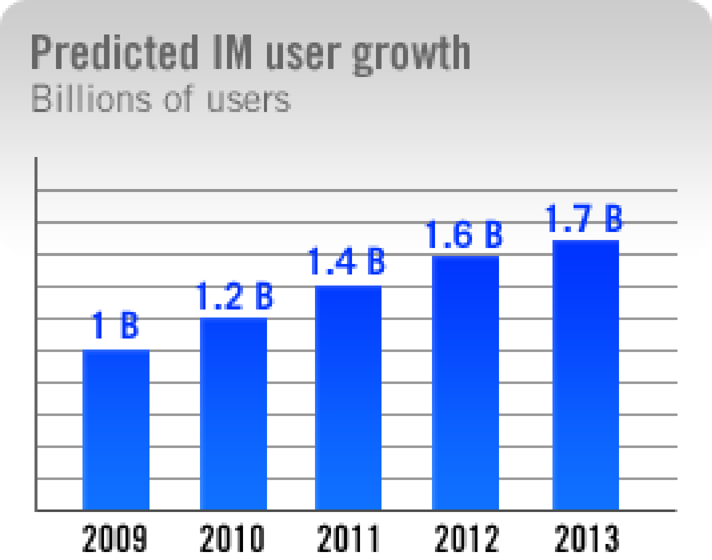


www.pingdom.com

Instant Messaging Facts



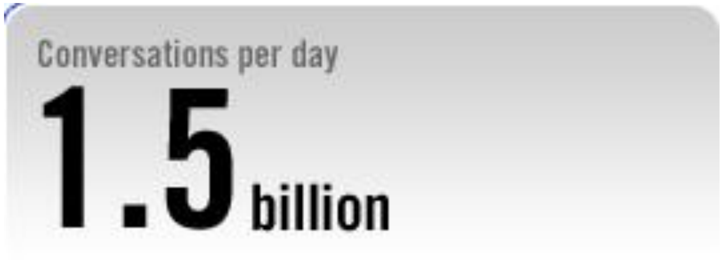
- ➢10000 US laws and regulations related to IM.
- ➢According to 2009 statistics:
- ➢ Around 47 billion IMs/day.
- ➢32% of IMs used by Enterprises.
- ➢53 IM messages/user daily.



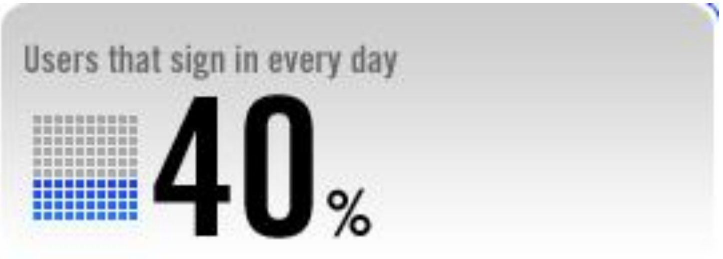
www.pingdom.com

Instant Messaging Facts

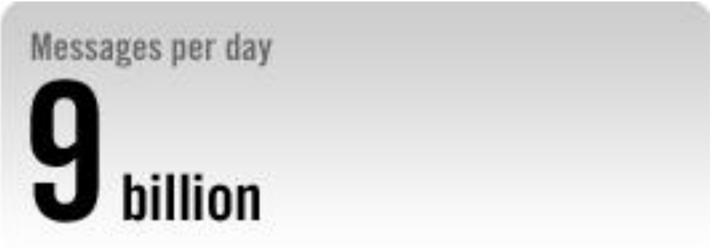
Live messenger


Conversations per day

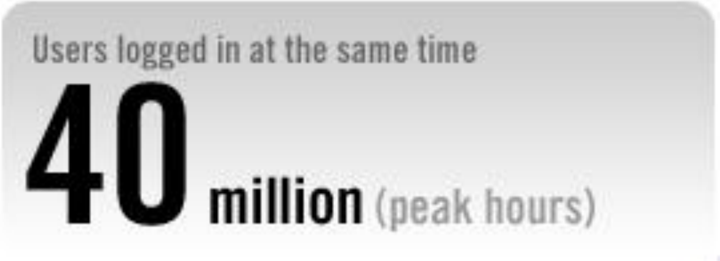
1.5 billion


Users that sign in every day

40%


Messages per day

9 billion

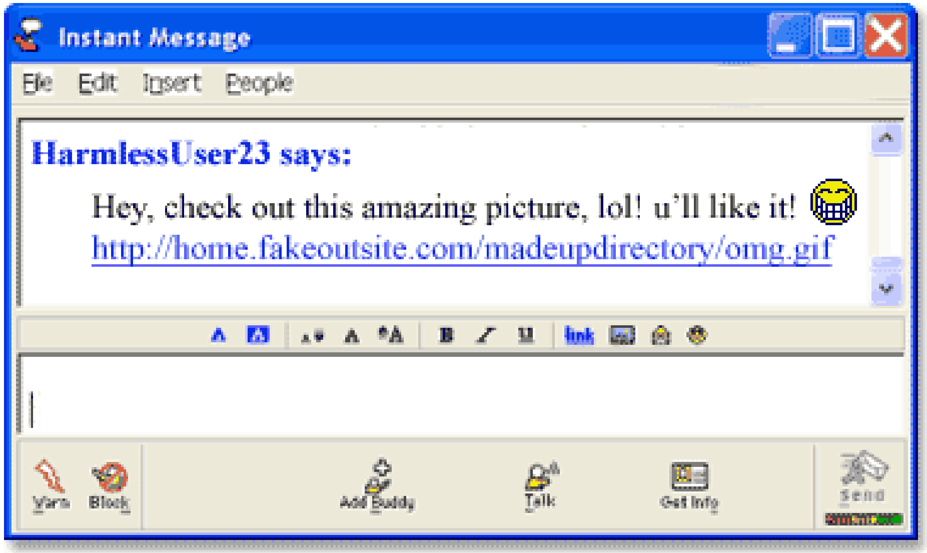

Users logged in at the same time

40 million (peak hours)

www.pingdom.com

Instant Messaging Facts

- IM is Convenient for Hackers.



HarmlessUser23 says:

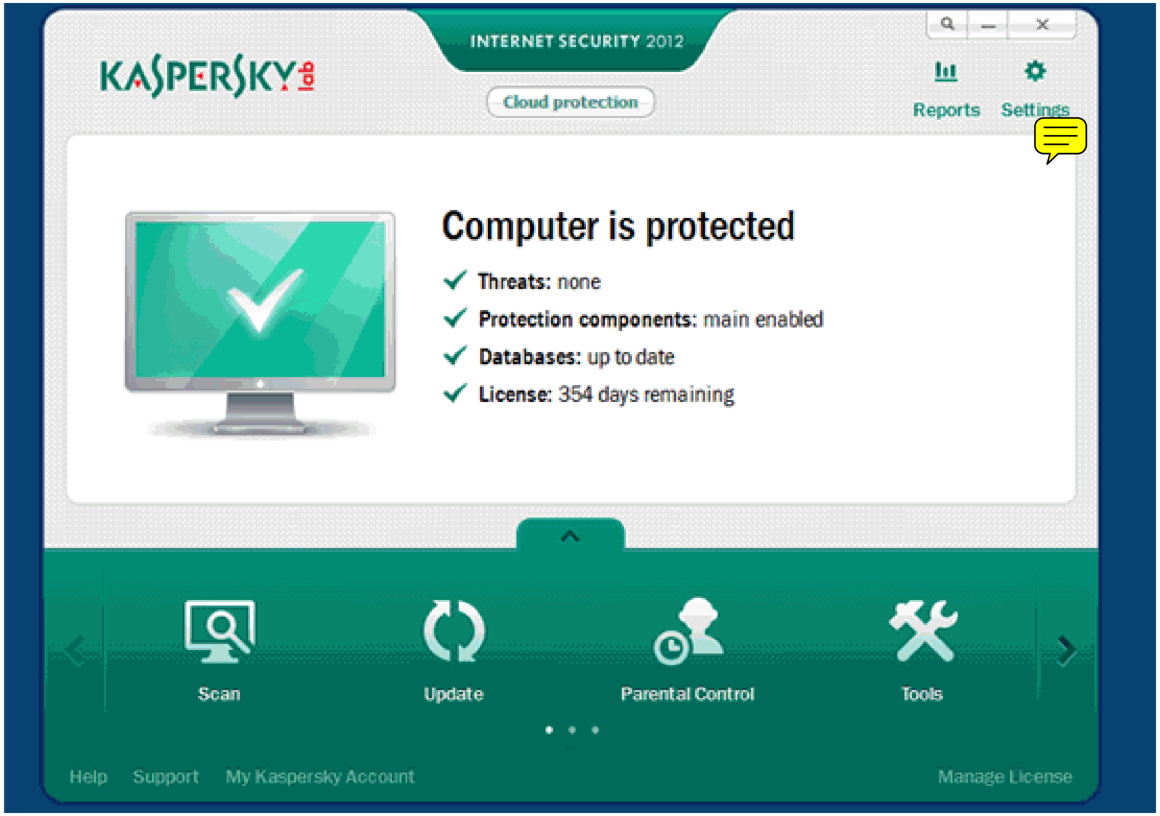
Hey, check out this amazing picture, lol! u'll like it!
http://home.fakeoutsite.com/madeupdirectory/omg.gif

- Think twice before clicking! IM messages like this one

Instant Messaging Facts

Some Anti-virus applications include some parental control on IM messages such as:

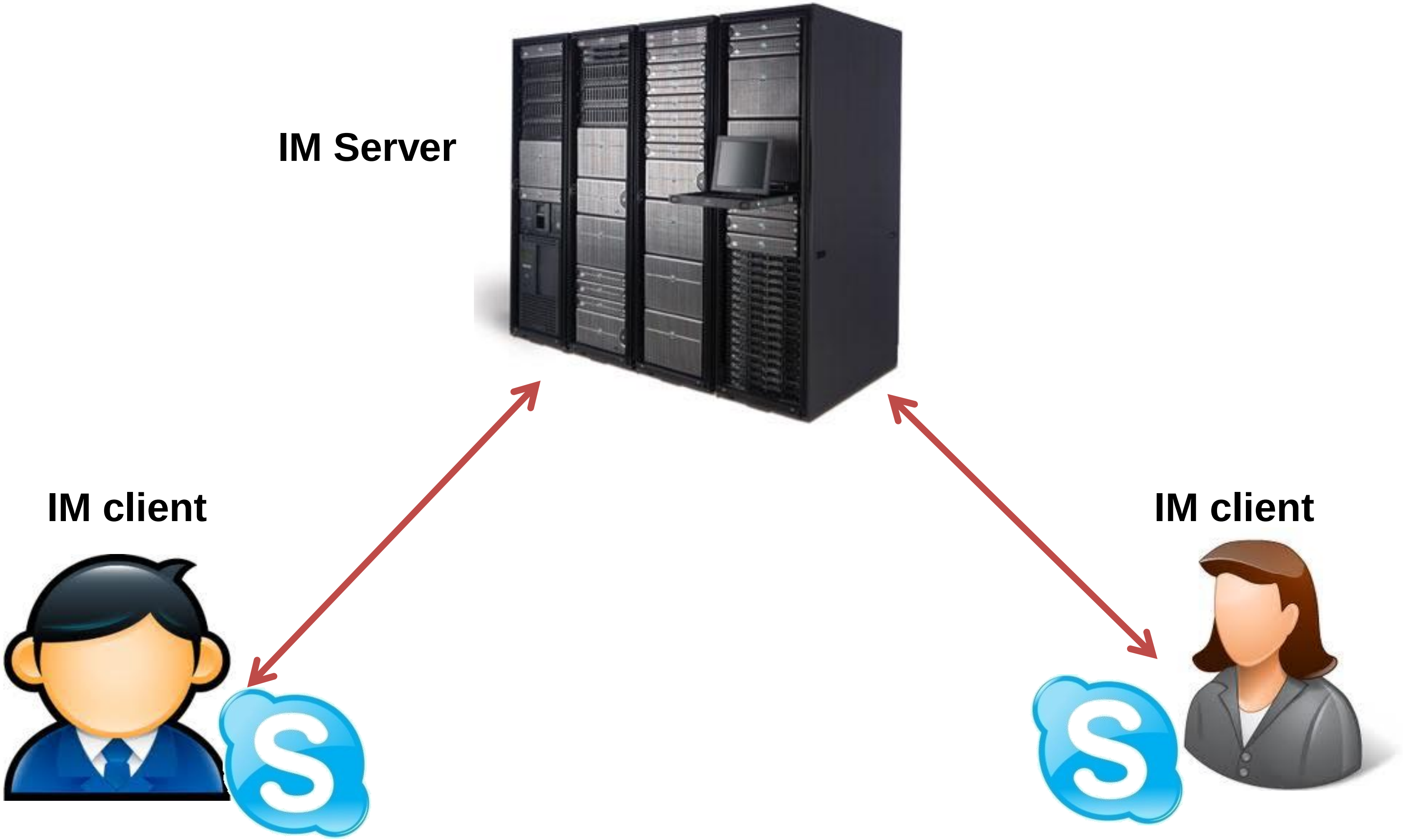
- Create lists of allowed and blocked contacts.
- Specify key words that all incoming messages will be checked for.
- Enter personal data prohibited to be sent.



Computer is protected

✓ Threats: none
✓ Protection components: main enabled
✓ Databases: up to date
✓ License: 354 days remaining

IM Architecture

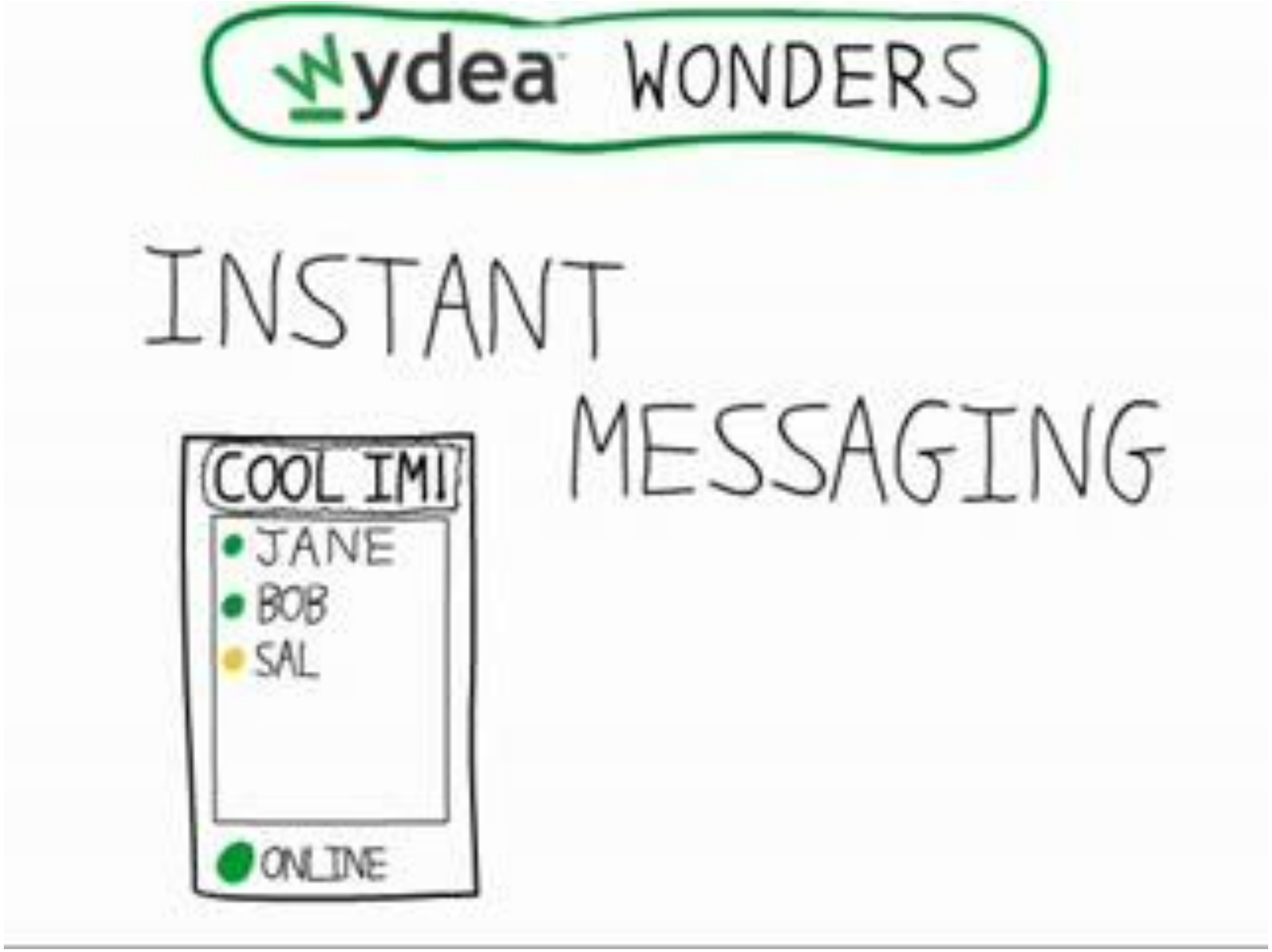


IM Server

IM client

IM client

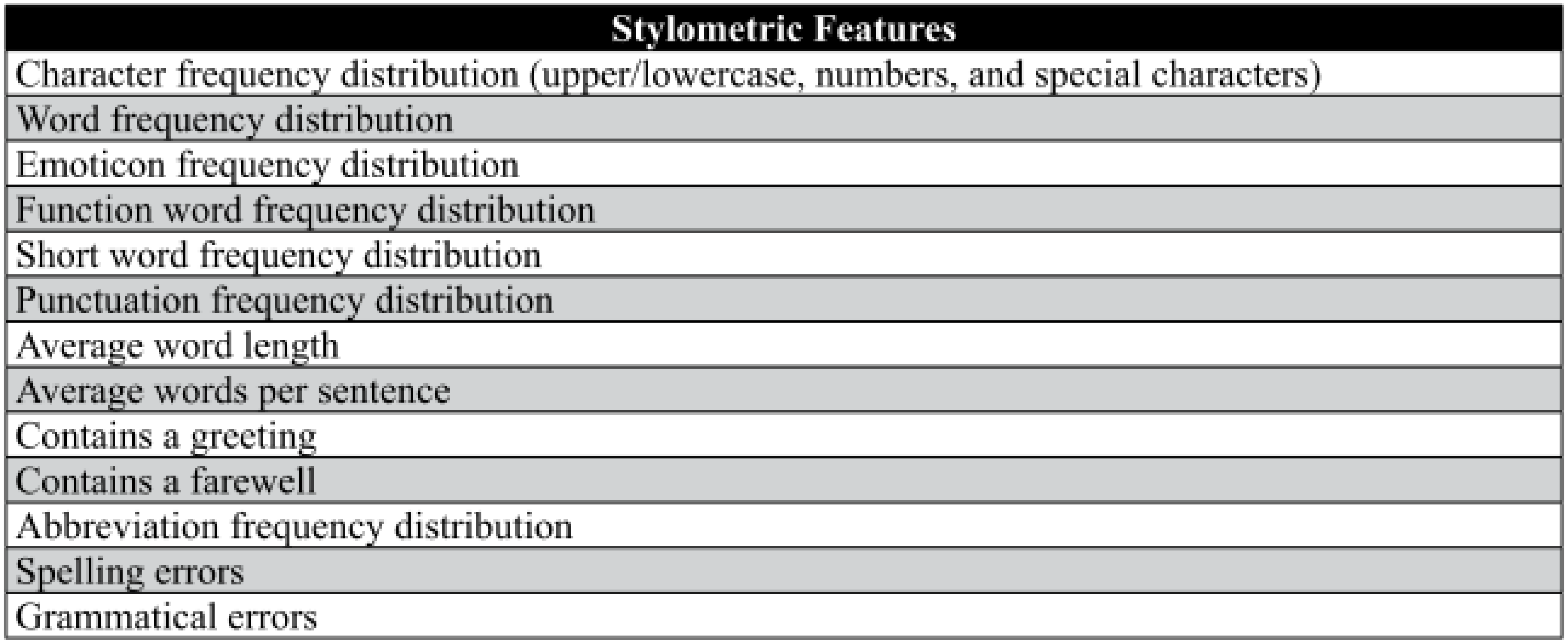
IM Architecture



Author Behavior Categorization

- Stylometric features: An author’s relatively constant set of characteristics for a large number of IM messages.
  - ➢Syntactic and structural layout traits.
  - ➢Patterns.
  - ➢Vocabulary usage.
  - ➢Unusual language usage.

Stylometric Features



Stylometric Features
Character frequency distribution (upper/lowercase, numbers, and special characters)
Word frequency distribution
Emoticon frequency distribution
Function word frequency distribution
Short word frequency distribution
Punctuation frequency distribution
Average word length
Average words per sentence
Contains a greeting
Contains a farewell
Abbreviation frequency distribution
Spelling errors
Grammatical errors

List of stylometric features may be used for IM author classification

Stylometric Features

Abbreviation	Sentence
1DR	I wonder
10Q	Thank you
LOL	laughing out loud
ROTFL	rolling on the floor laughing
RU	are you
4	for
HW	Homework
4EAE	Forever and ever

Experiments and Results

Data Description:

- Gaim and Adium clients conversations log.
- Conversation Format:
  - [timestamp] [user name:] [message]
- Example:
  - (14:19:29) User1: hey, what time is the meeting today?
  - (14:19:35) User2: It is at 11AM…are you going?
  - (14:19:39) User1: yeah, I will be there, it sounds very interesting! :) :)

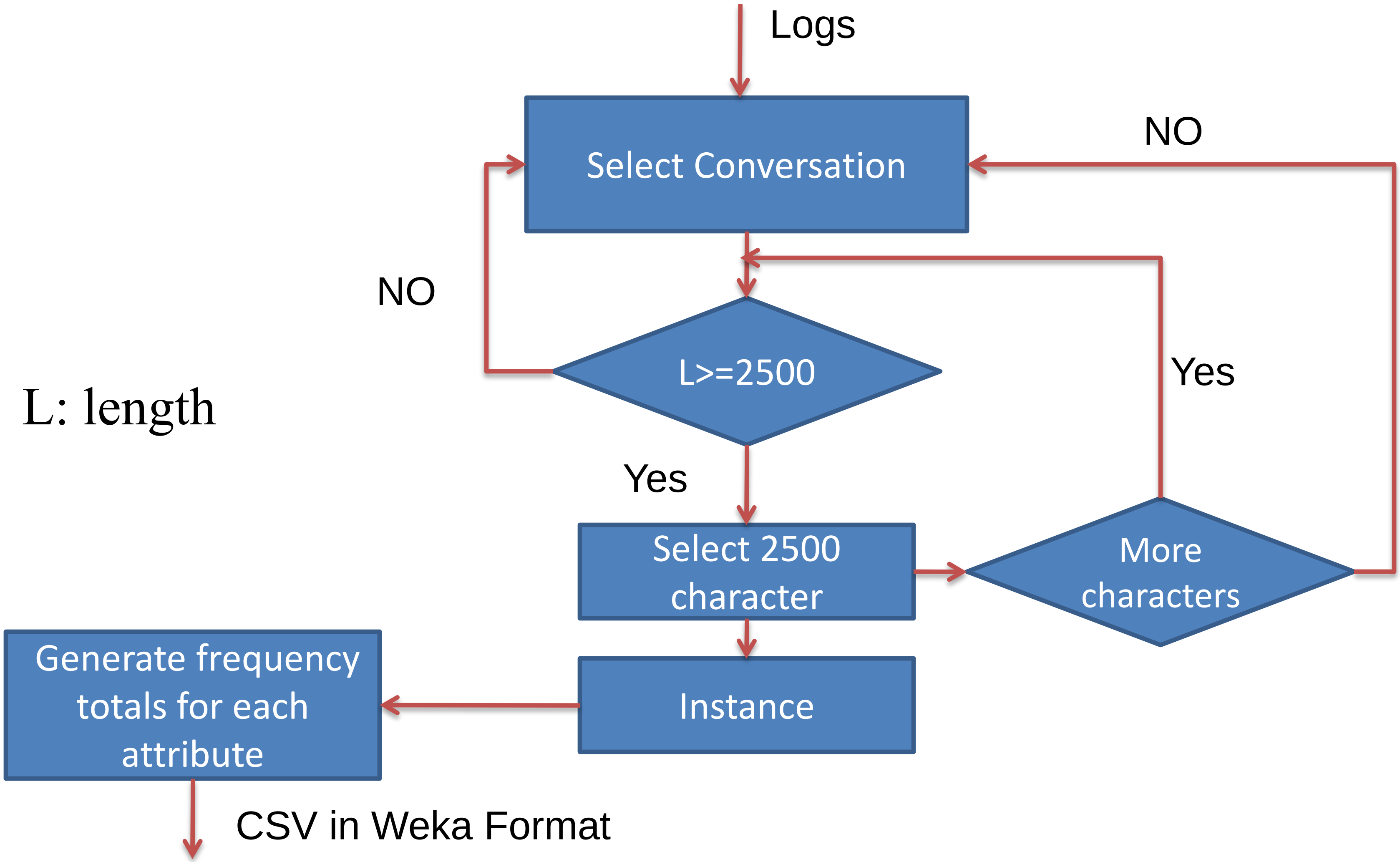
Experiments and Results

Data preparation for analysis:

1. All entries that did not belong to specific user are removed.
2. Timestamp and username are removed too.

- Example of prepared data for user 1:
  - ➢ hey, what time is the meeting today?
  - ➢ yeah, I will be there, it sounds very interesting! :) :)

Experiments and Results

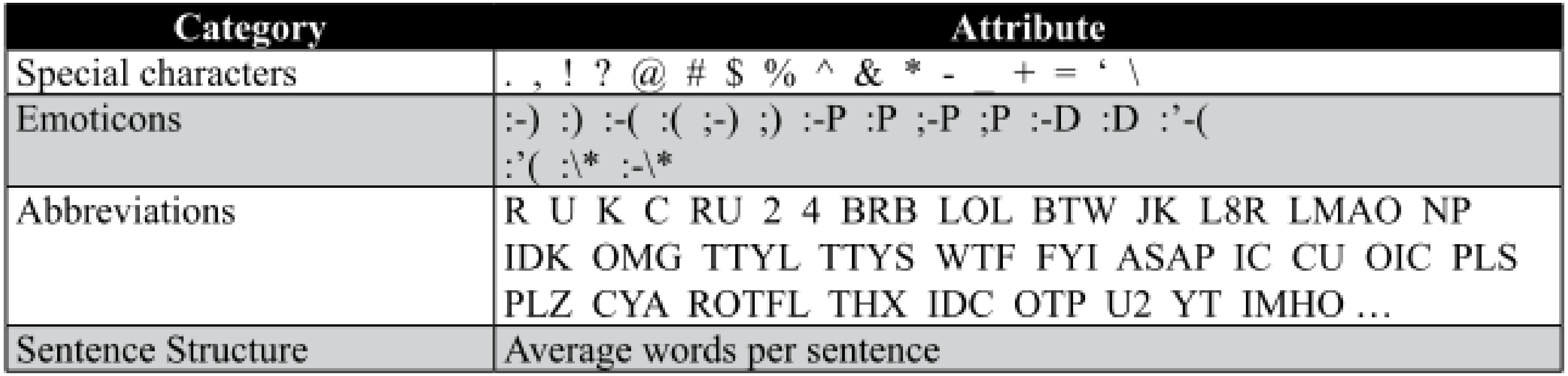


L: length

Experiments and Results

Data used in research:

- ➢ logs for four users (User 1,2,3 and 4).
- ➢ 35 instances of 2500 characters for each user.
- ➢ 69 numeric attributes

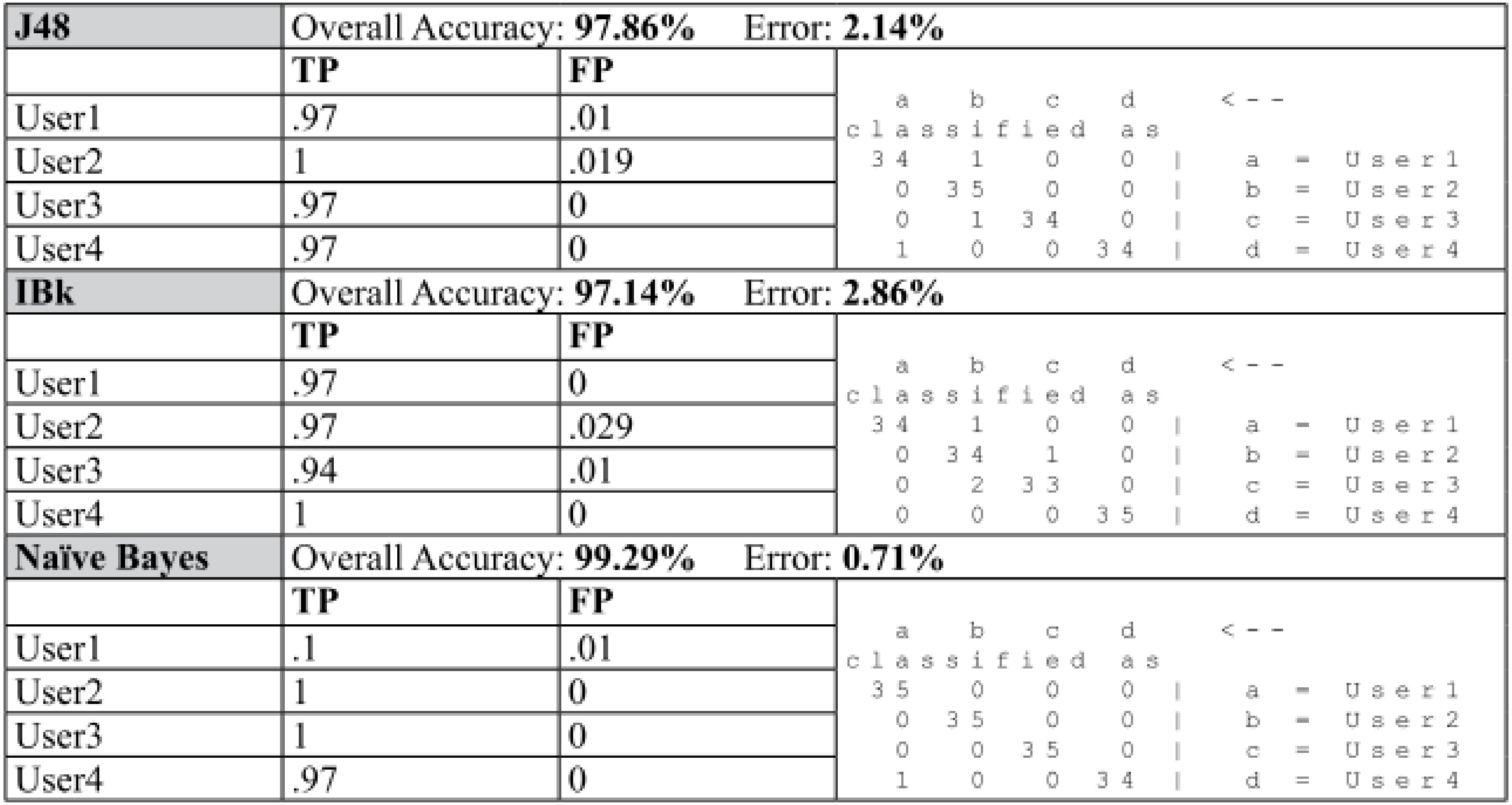


Category	Attribute	
Special characters	. , ! ? @ # $ % ^ & * - _ + = ‘ \	17
Emoticons	:-) :) :-(:(;-) ;) :-P :P ;-P ;P :-D :D :’-(:’(:* :-*	16
Abbreviations	R U K C RU 2 4 BRB LOL BTW JK L8R LMAO NP IDK OMG TTYL TTYS WTF FYI ASAP IC CU OIC PLS PLZ CYA ROTFL THX IDC OTP U2 YT IMHO …	35
Sentence Structure	Average words per sentence	1

Experiments and Results

- Weka data mining are used for classification.
- Classifiers used:
  - J48 decision tree.
  - IBk nearest neighbor.
  - Naïve Bayes classifiers.

Experiments and Results



J48	Overall Accuracy: **97.86%** Error: **2.14%**	
	TP	**FP**
User1	.97	.01
User2	1	.019
User3	.97	0
User4	.97	0

```
  a   b   c   d    <--
classified  as
 34   1   0   0  |   a = User1
  0  35   0   0  |   b = User2
  0   1  34   0  |   c = User3
  1   0   0  34  |   d = User4
```

IBk	Overall Accuracy: **97.14%** Error: **2.86%**	
	TP	**FP**
User1	.97	0
User2	.97	.029
User3	.94	.01
User4	1	0

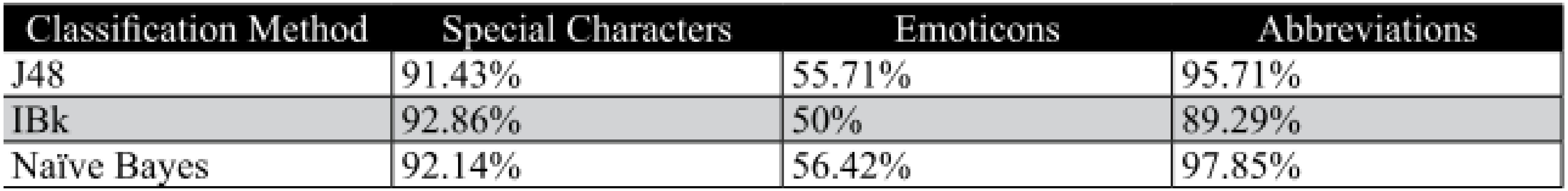
```
  a   b   c   d    <--
classified  as
 34   1   0   0  |   a = User1
  0  34   1   0  |   b = User2
  0   2  33   0  |   c = User3
  0   0   0  35  |   d = User4
```

Naïve Bayes	Overall Accuracy: **99.29%** Error: **0.71%**	
	TP	**FP**
User1	.1	.01
User2	1	0
User3	1	0
User4	.97	0

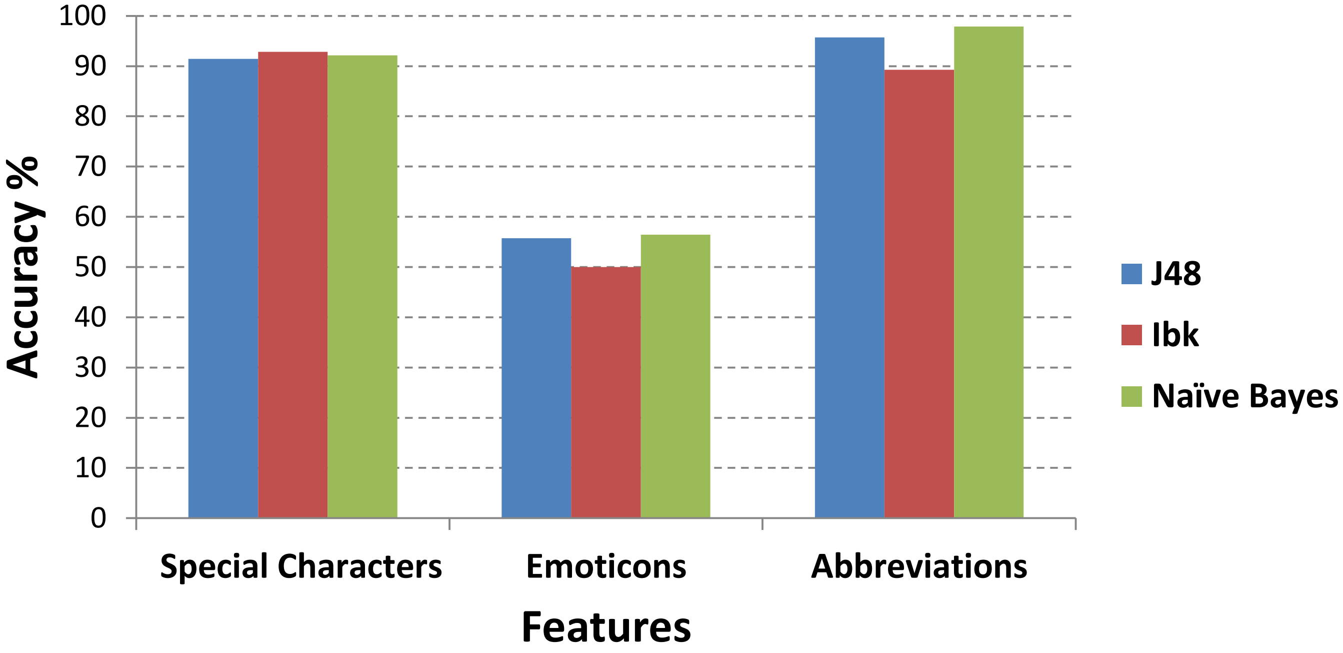
```
  a   b   c   d    <--
classified  as
 35   0   0   0  |   a = User1
  0  35   0   0  |   b = User2
  0   0  35   0  |   c = User3
  1   0   0  34  |   d = User4
```

IM Data Classification Results

Experiments and Results



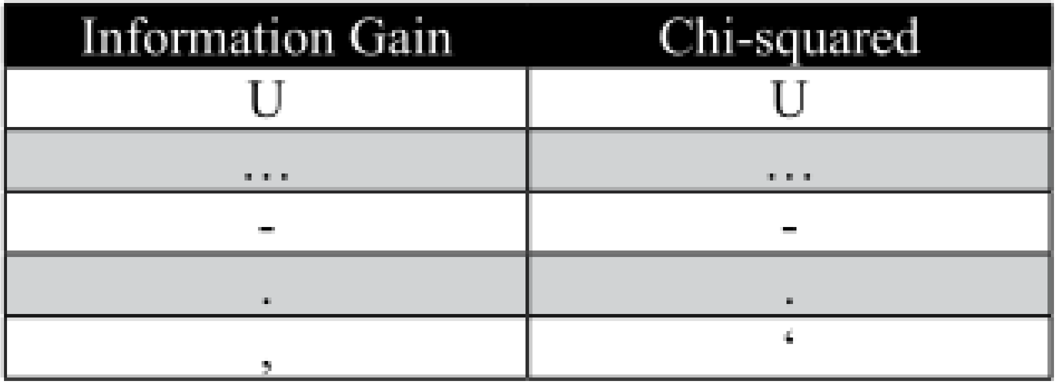
Classification Method	Special Characters	Emoticons	Abbreviations
J48	91.43%	55.71%	95.71%
IBk	92.86%	50%	89.29%
Naïve Bayes	92.14%	56.42%	97.85%



Classification Accuracy Results for individual Attribute Categories

Experiments and Results

- Attribute selection was used to rank the strongest attributes in identifying process.

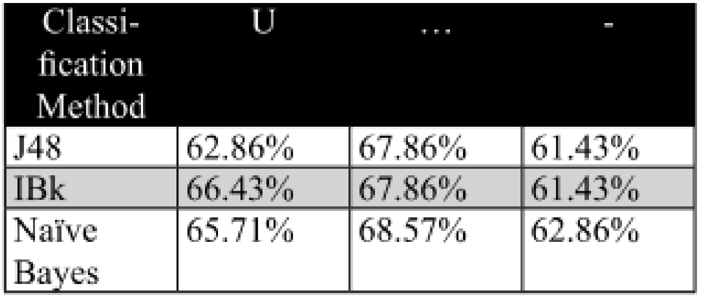


Information Gain	Chi-squared
U	U
...	...
-	-
.	.
,	،

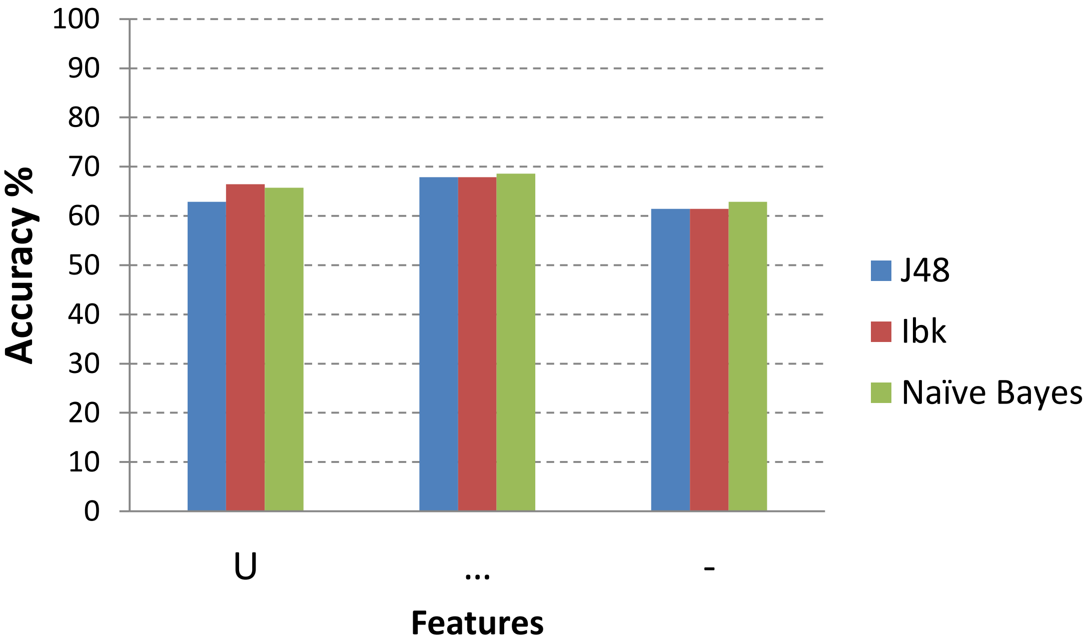
Attribute Selection

Experiments and Results

- The top 3 individual attributes (U, three dots, and the hyphen) were tested individually.



Classi-fication Method	U	…	-
J48	62.86%	67.86%	61.43%
IBk	66.43%	67.86%	61.43%
Naïve Bayes	65.71%	68.57%	62.86%



Results Discussion

- The best discriminators:
  - Abbreviations (97.85% accuracy).
  - Special characters (92.86% accuracy).
- The Naïve Bayes performed (97.85%) with the abbreviations only.
- J48 and IBk classifiers performed (97.86%) and (97.14%) when all attributes combined.

Results Discussion

- The strongest identifying attributes are U, Three dots the hyphen.
- None of the individual attributes were strong enough to determine author identification.
- Naïve Bayes classification provided the best results (99.29% accuracy) when all attributes combined.

Summary

- Recently, many users use IM for online conversations.
- This area is not explored well.
- several concerns involving the use of IM systems (man-in-the-middle attacks).
- This paper uses data mining of IM communications for authorship identification.

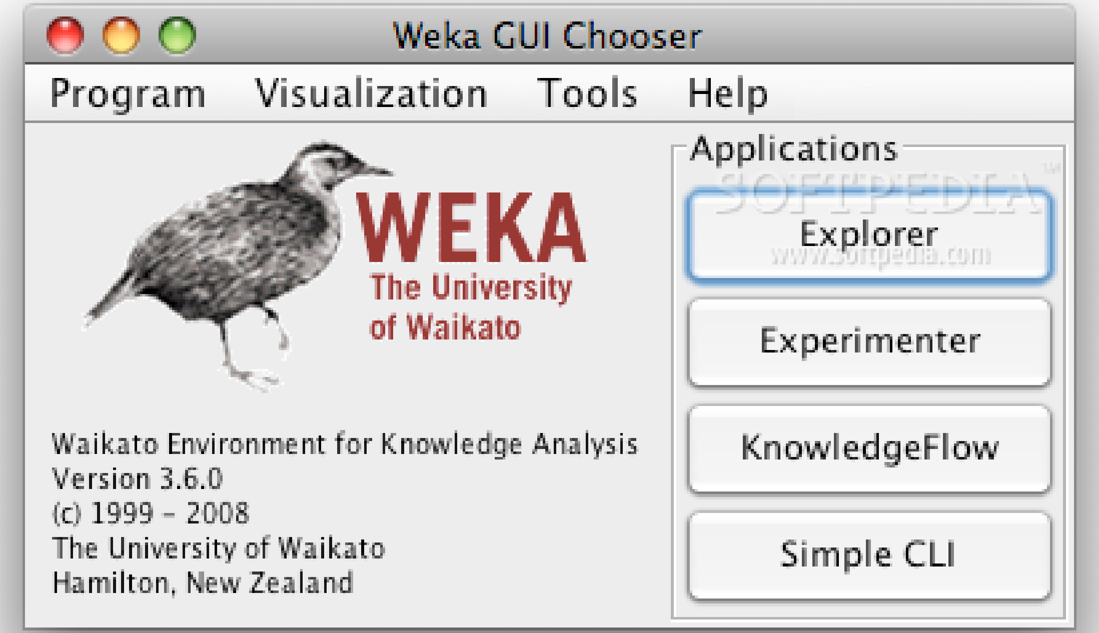
Summary

- Classification features based on authors various behaviors.
- Results show that Naïve Bayes is highly accurate (> 99% accuracy).
- Identification of the strongest behavior characteristics.

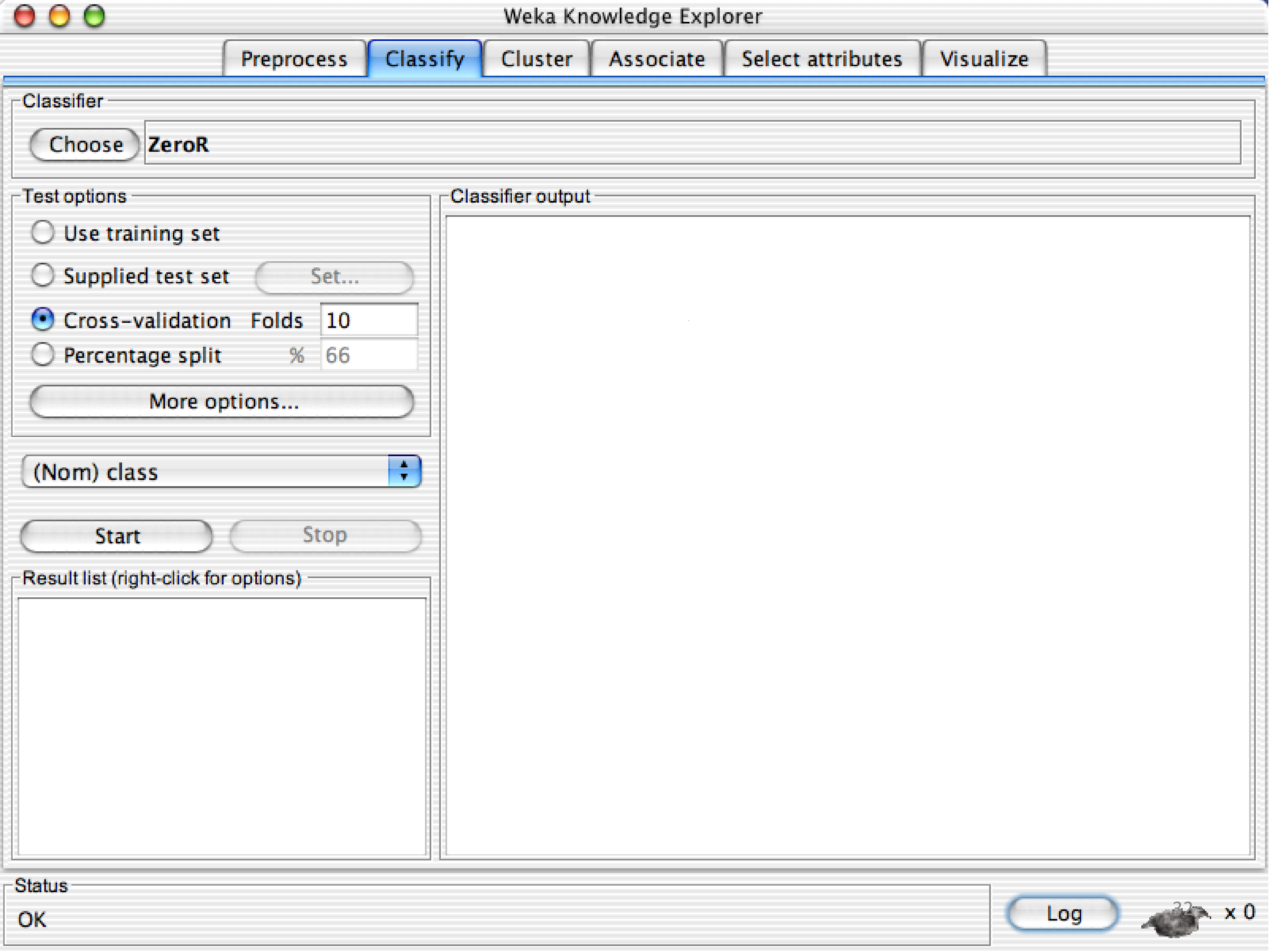
Future Work

- Increase the numbers of users.
- Increase the numbers of attributes.
- Varied the numbers of characters in an instance (determine the minimum size necessary for high accuracy).
- Include other stylometric measures.

Weka Classifier Selection Steps



Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand



Weka Knowledge Explorer

Classifier

ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

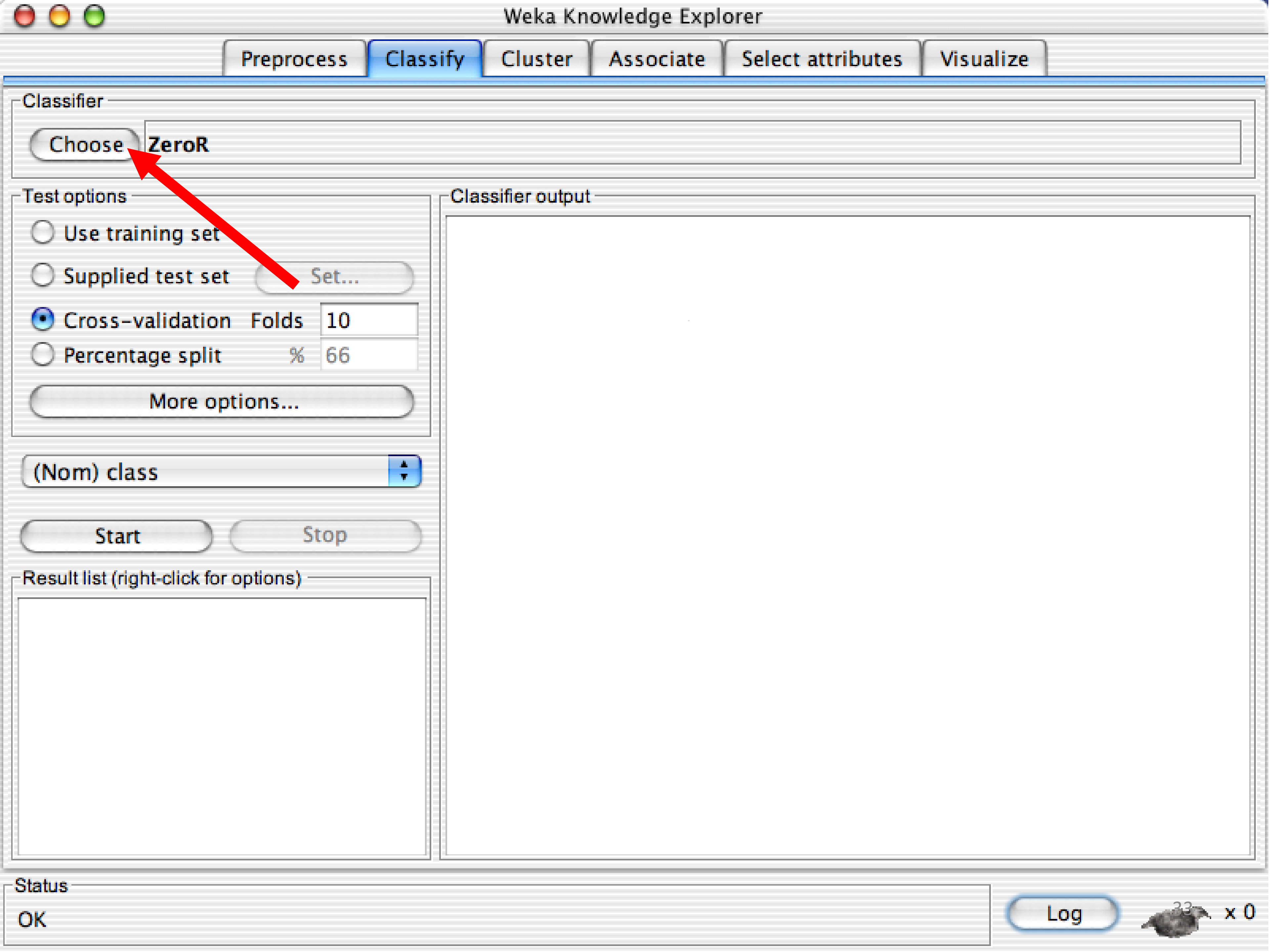
(Nom) class

Result list (right-click for options)

Classifier output

Status

OK



Classifier

ZeroR

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

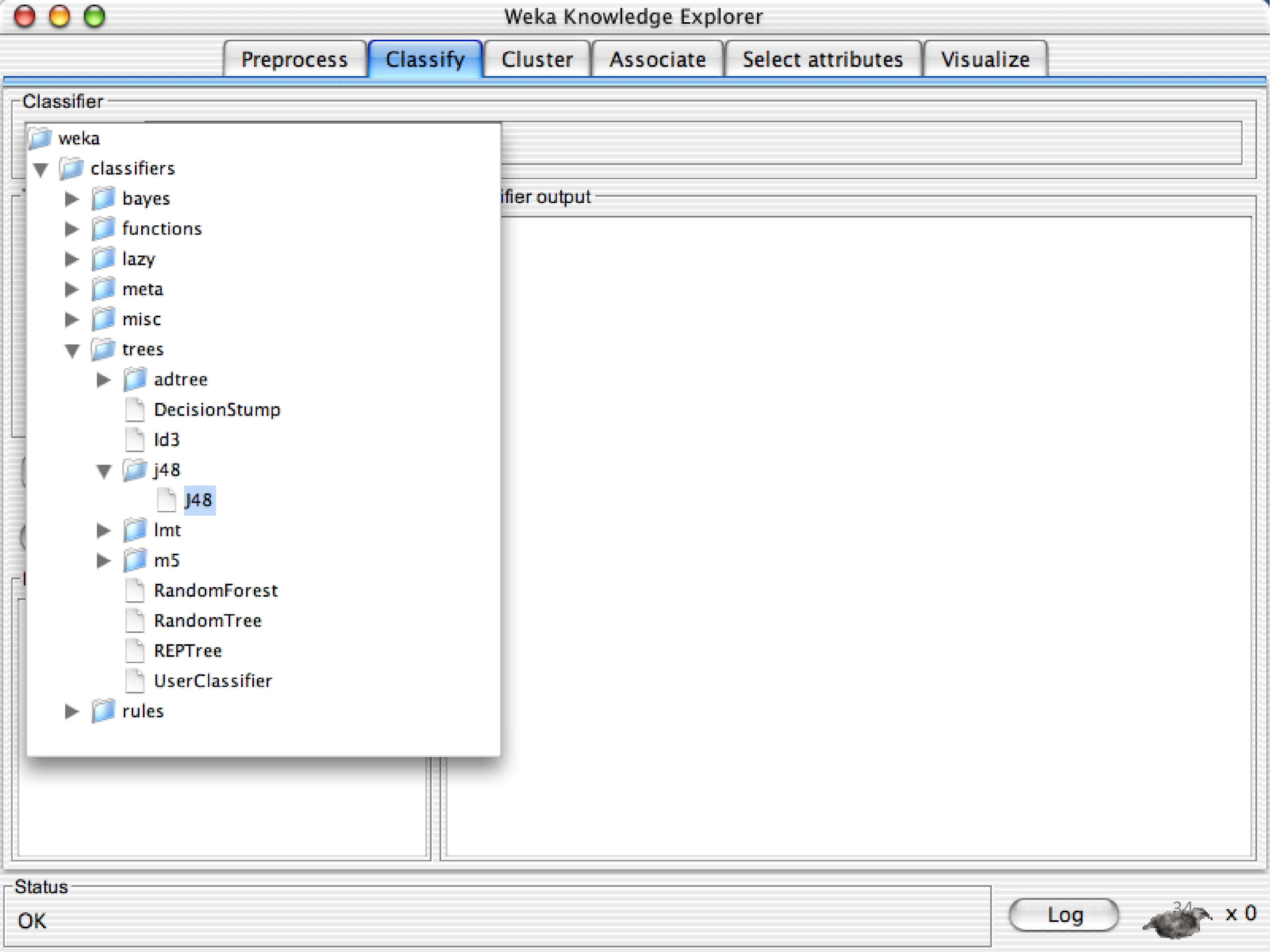
(Nom) class

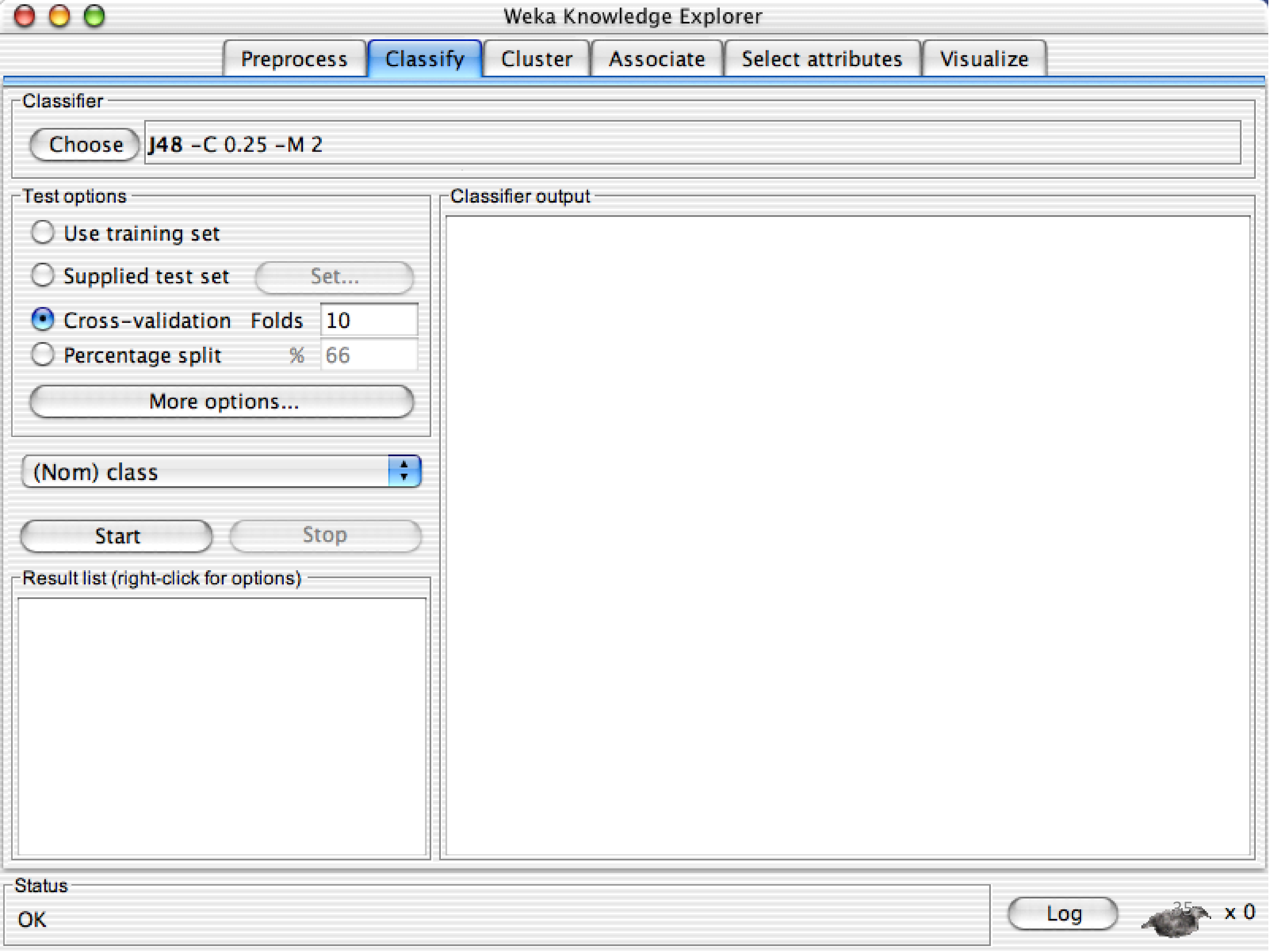
Result list (right-click for options)

Classifier output

Status

OK


Weka Knowledge Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
weka
classifiers
bayes
functions
lazy
meta
misc
trees
adtree
DecisionStump
Id3
j48
J48
lmt
m5
RandomForest
RandomTree
REPTree
UserClassifier
rules
ifier output
Status
OK
Log
x 0



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

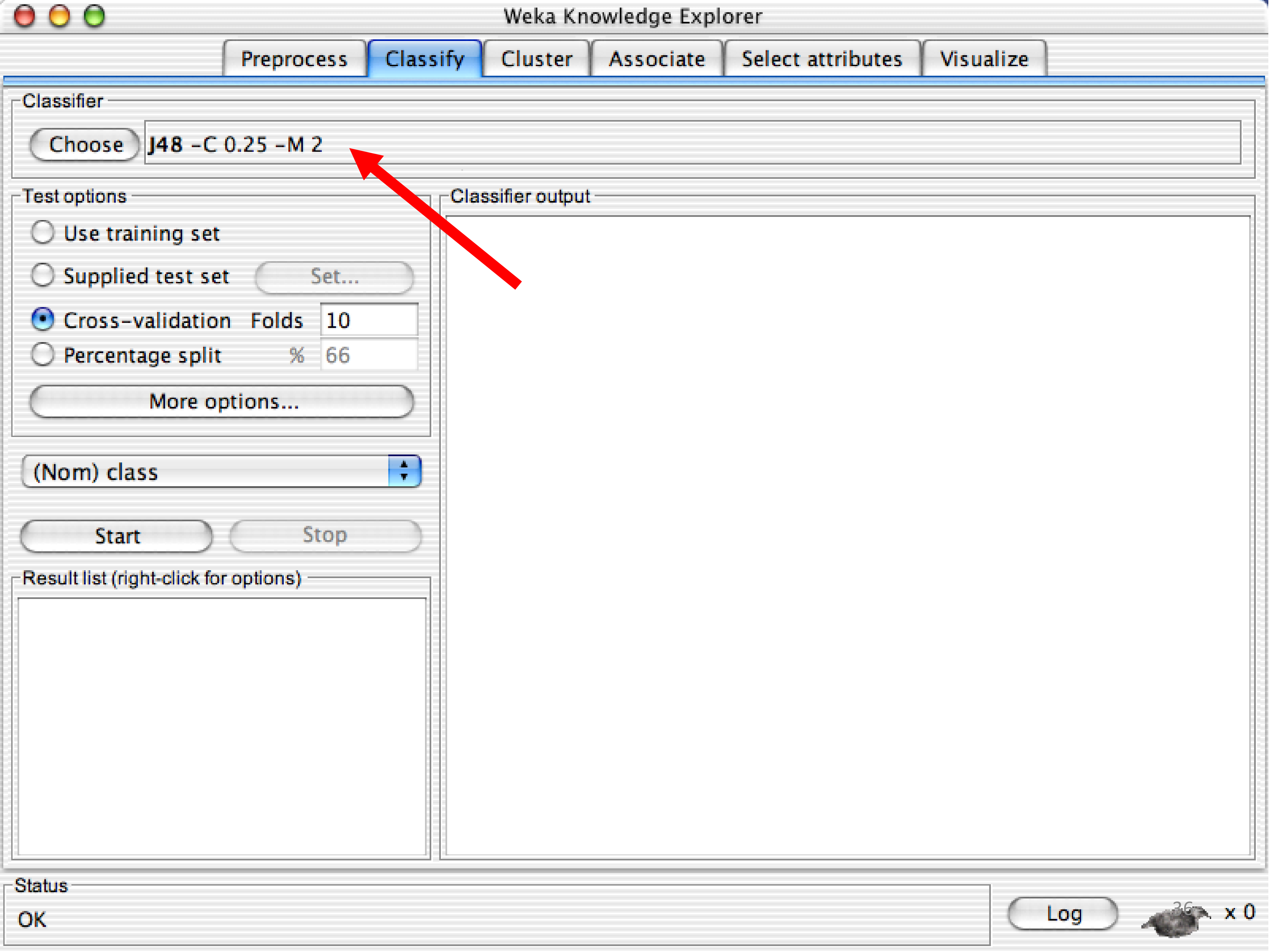
○ Percentage split % 66

(Nom) class

Status

OK

x 0



Weka Knowledge Explorer

Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

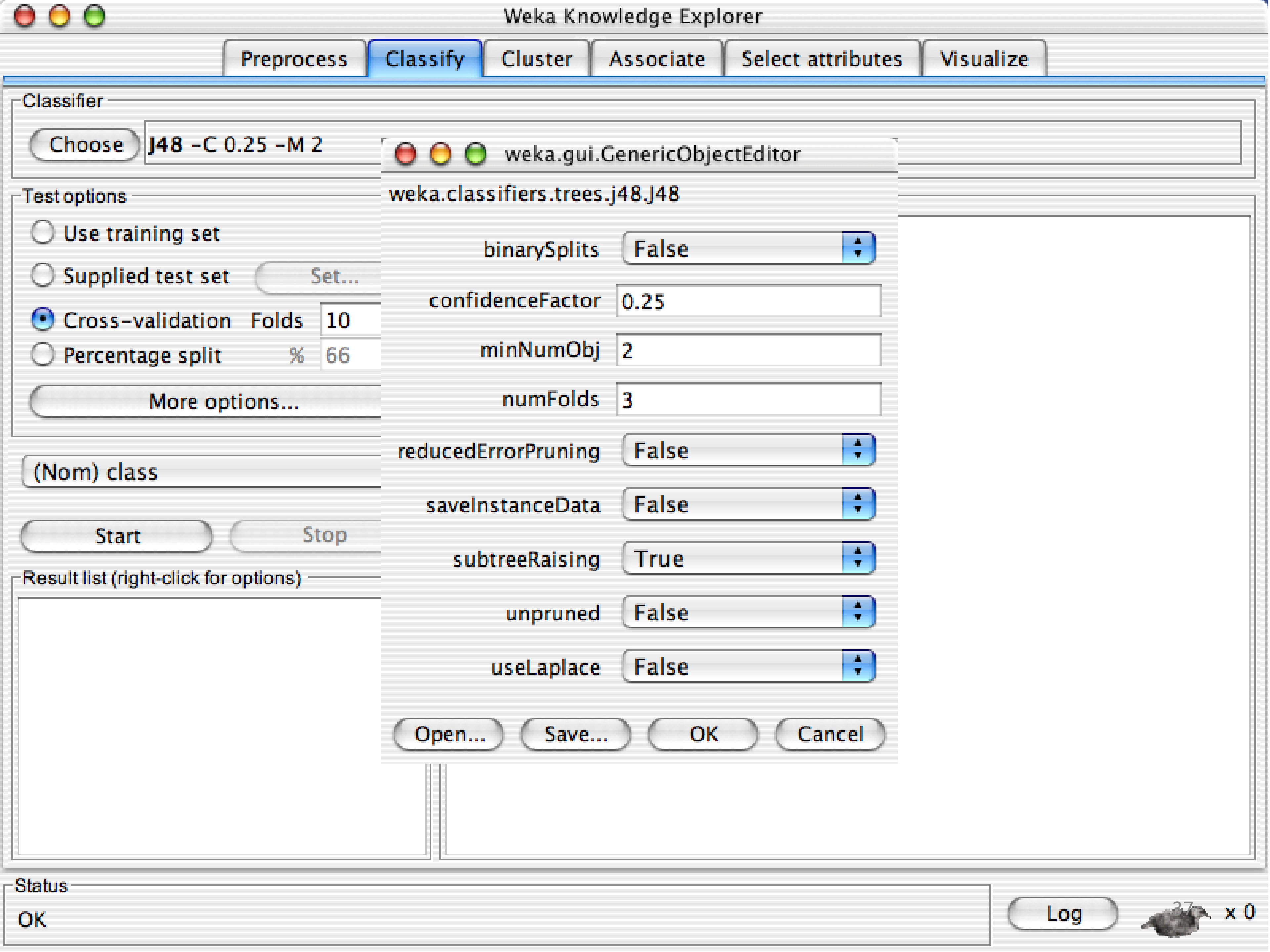
(Nom) class

Result list (right-click for options)

Classifier output

Status

OK



Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) class

Result list (right-click for options)

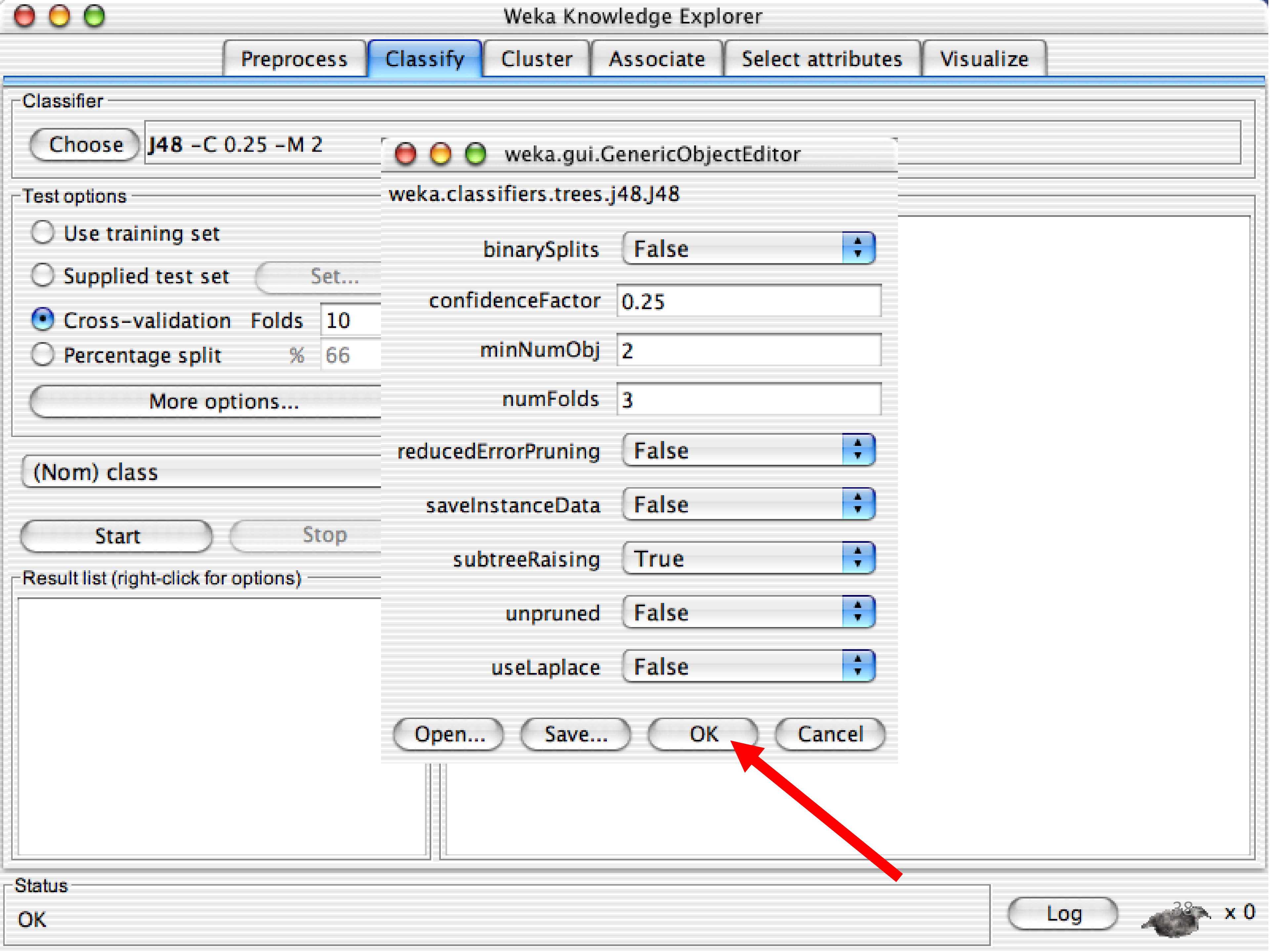
Status

OK

weka.gui.GenericObjectEditor

weka.classifiers.trees.j48.J48

binarySplits	False
confidenceFactor	0.25
minNumObj	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
subtreeRaising	True
unpruned	False
useLaplace	False



J48 -C 0.25 -M 2

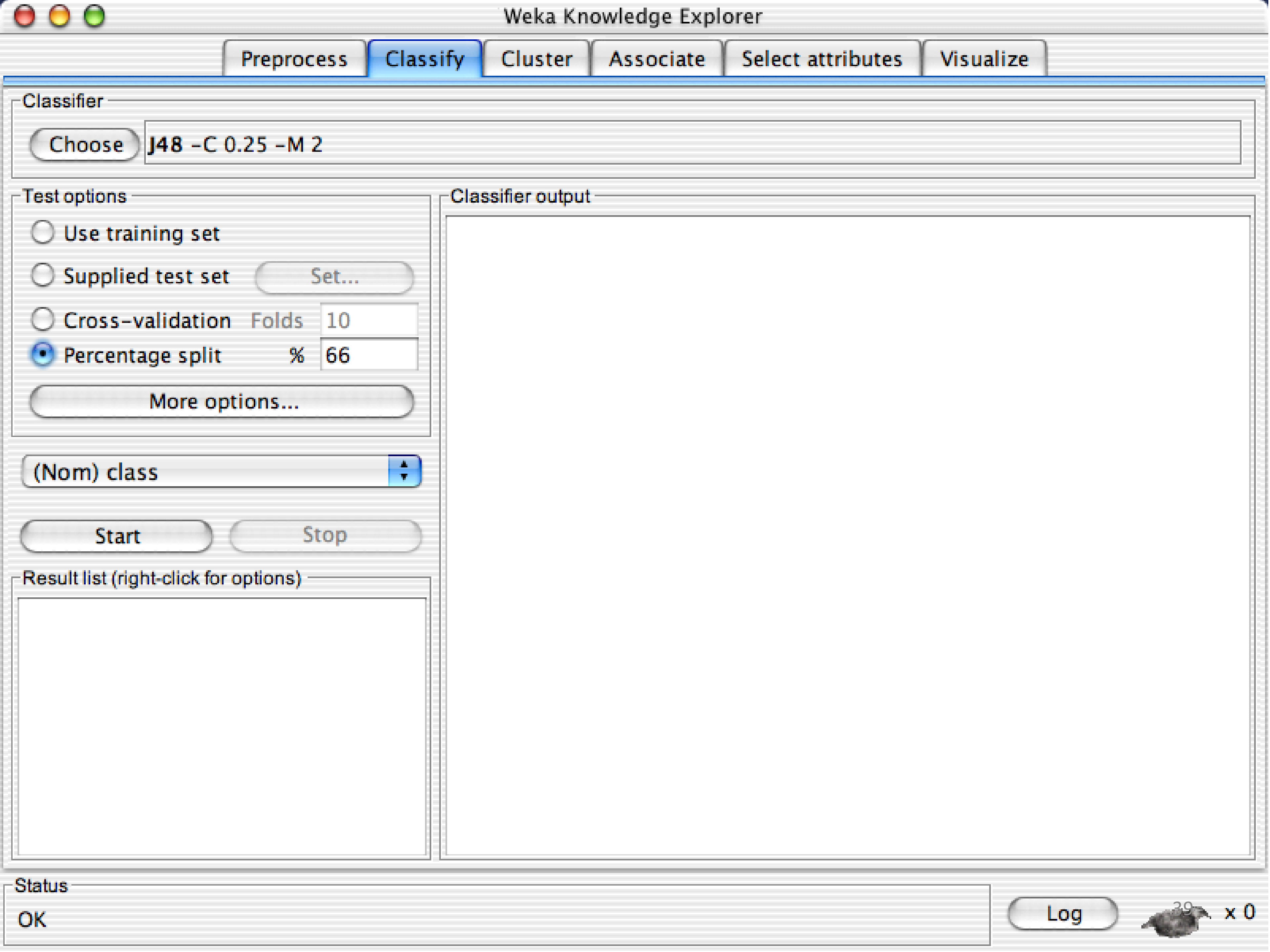
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

Status
OK

weka.classifiers.trees.j48.J48

binarySplits	False
confidenceFactor	0.25
minNumObj	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
subtreeRaising	True
unpruned	False
useLaplace	False



Weka Knowledge Explorer

Classifier: J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

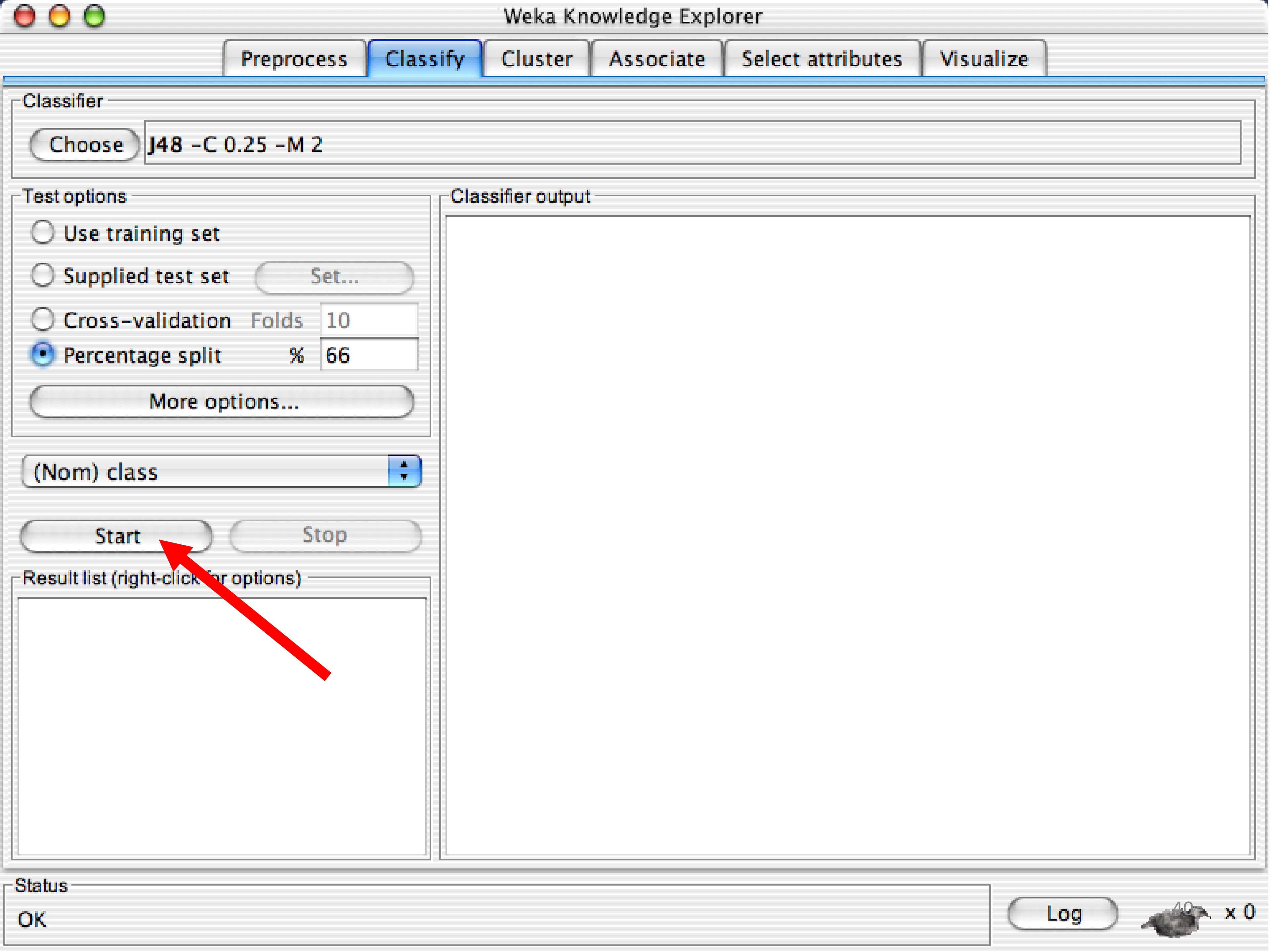
○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

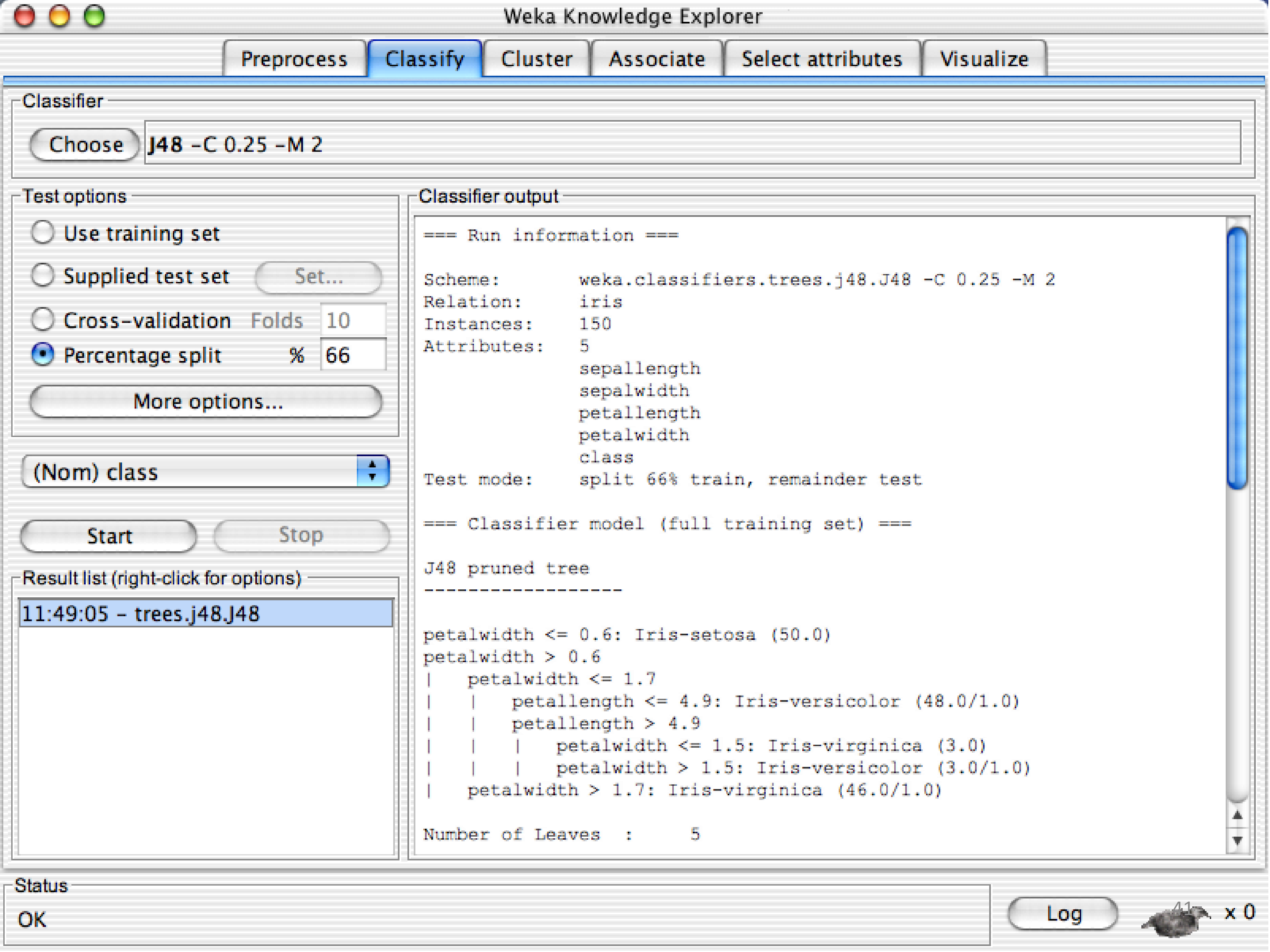
(Nom) class

Result list (right-click for options)

Classifier output

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

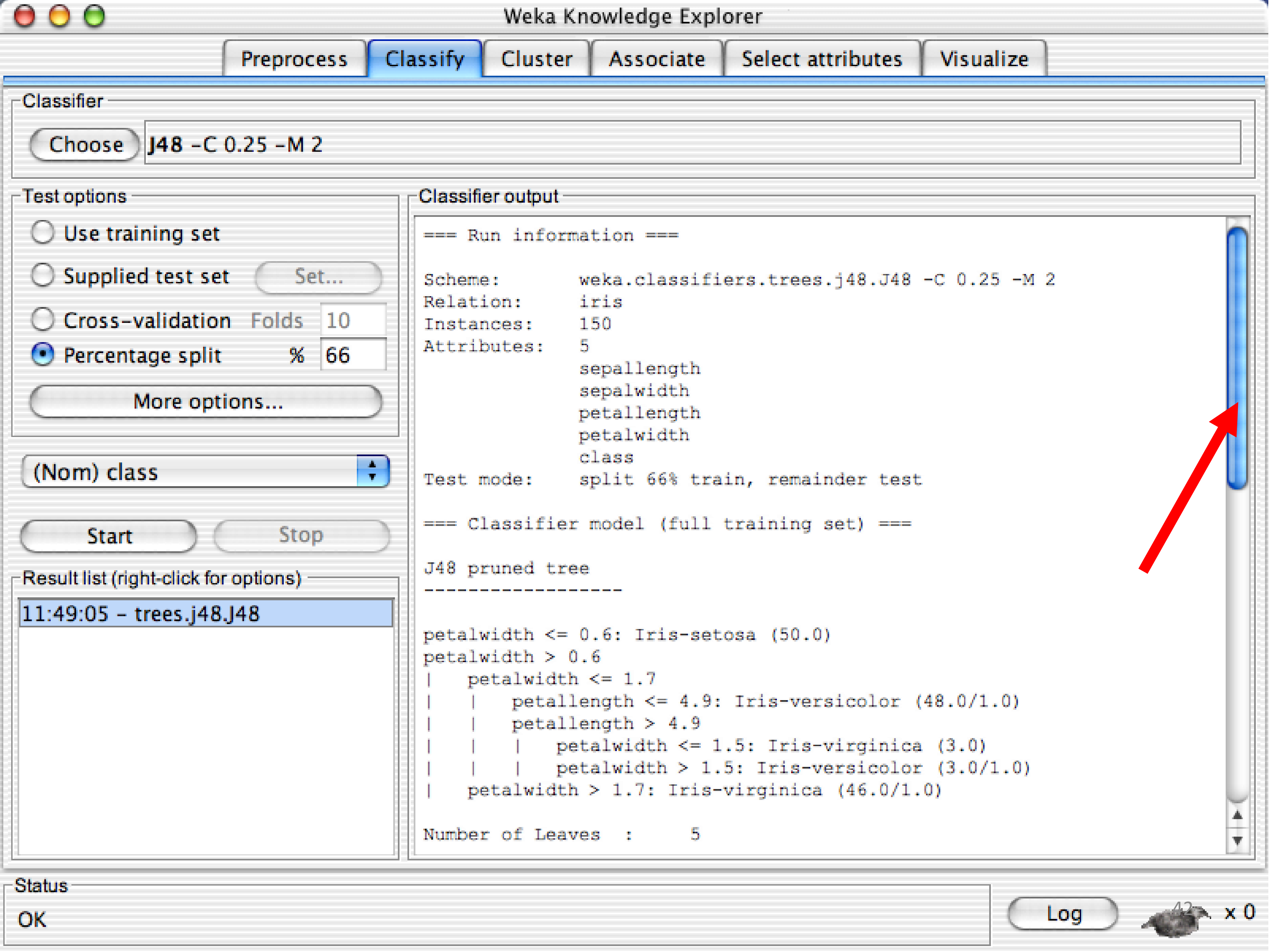
11:49:05 - trees.j48.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    split 66% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5
```

Status

OK



Classifier

J48 -C 0.25 -M 2

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) class

Result list (right-click for options)

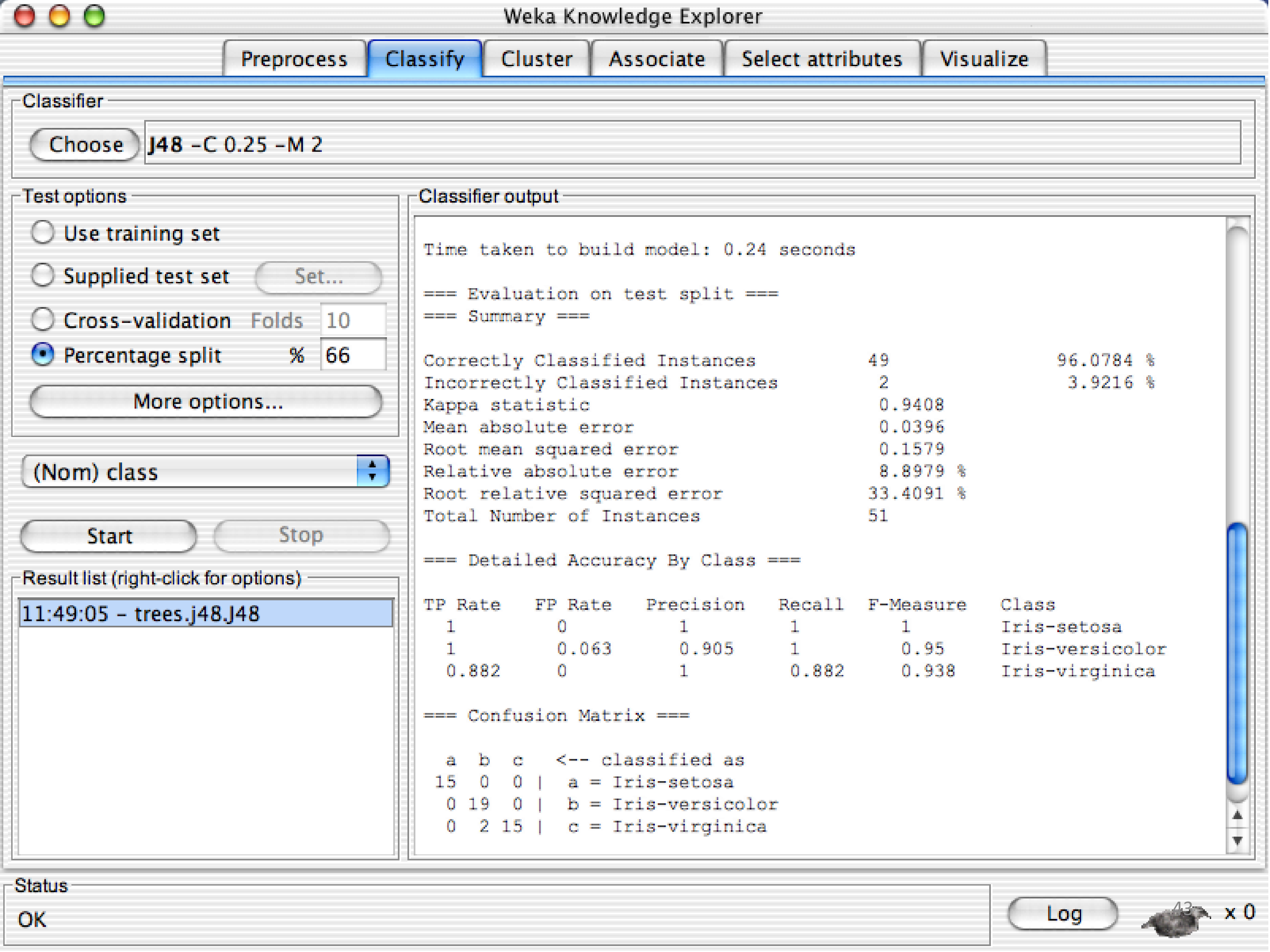
11:49:05 - trees.j48.J48

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.trees.j48.J48 -C 0.25 -M 2
Relation:     iris
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:    split 66% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree
------------------

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5
```

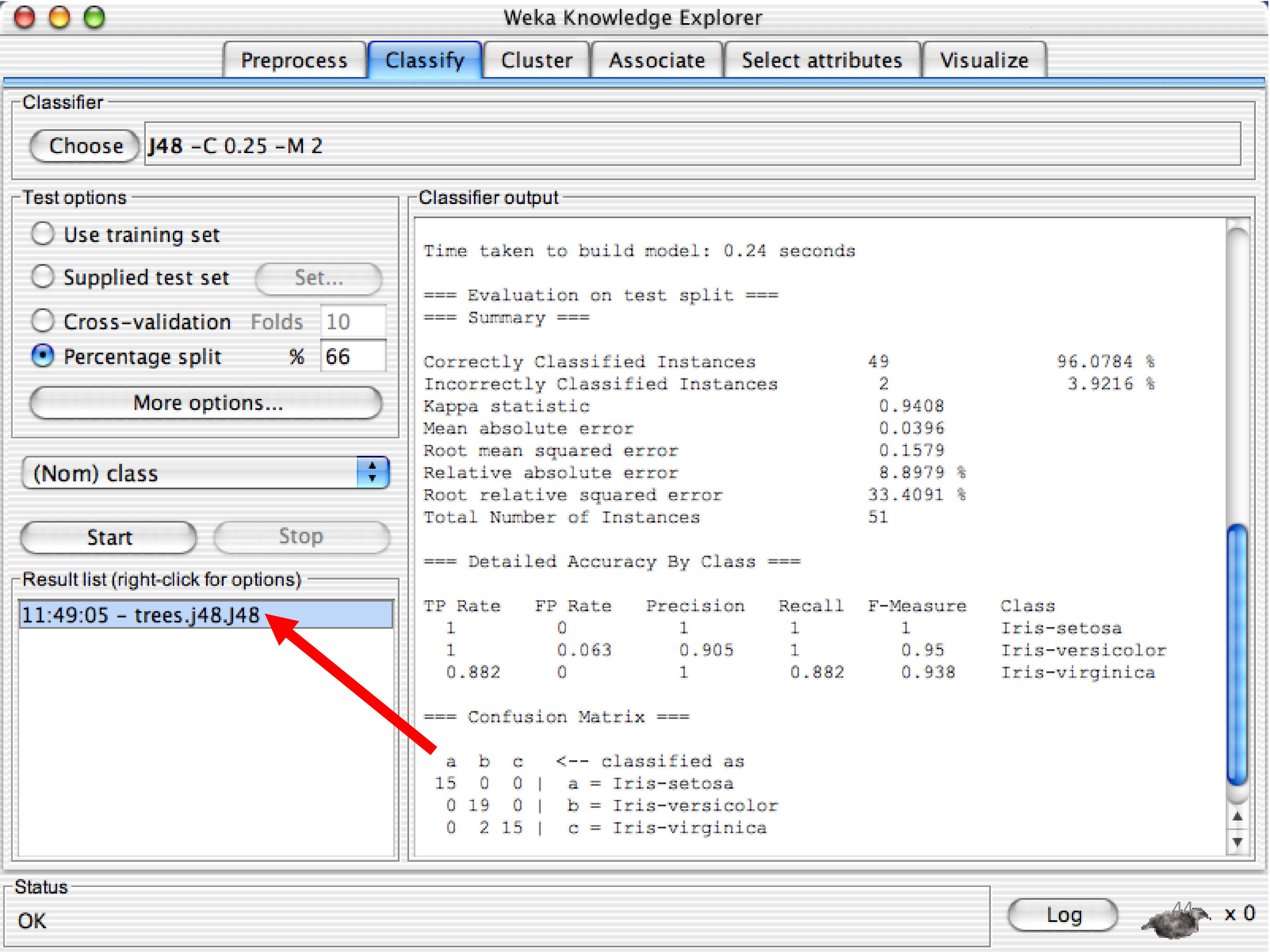
Status

OK



J48 -C 0.25 -M 2

```
Time taken to build model: 0.24 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0          1         1         1        Iris-setosa
  1         0.063      0.905     1         0.95     Iris-versicolor
  0.882     0          1         0.882     0.938    Iris-virginica

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  0  0 |  a = Iris-setosa
  0 19  0 |  b = Iris-versicolor
  0  2 15 |  c = Iris-virginica
```



Classifier

J48 -C 0.25 -M 2

Test options

- ○ Use training set
- ○ Supplied test set
- ○ Cross-validation Folds 10
- ● Percentage split % 66

(Nom) class

Result list (right-click for options)

11:49:05 - trees.j48.J48

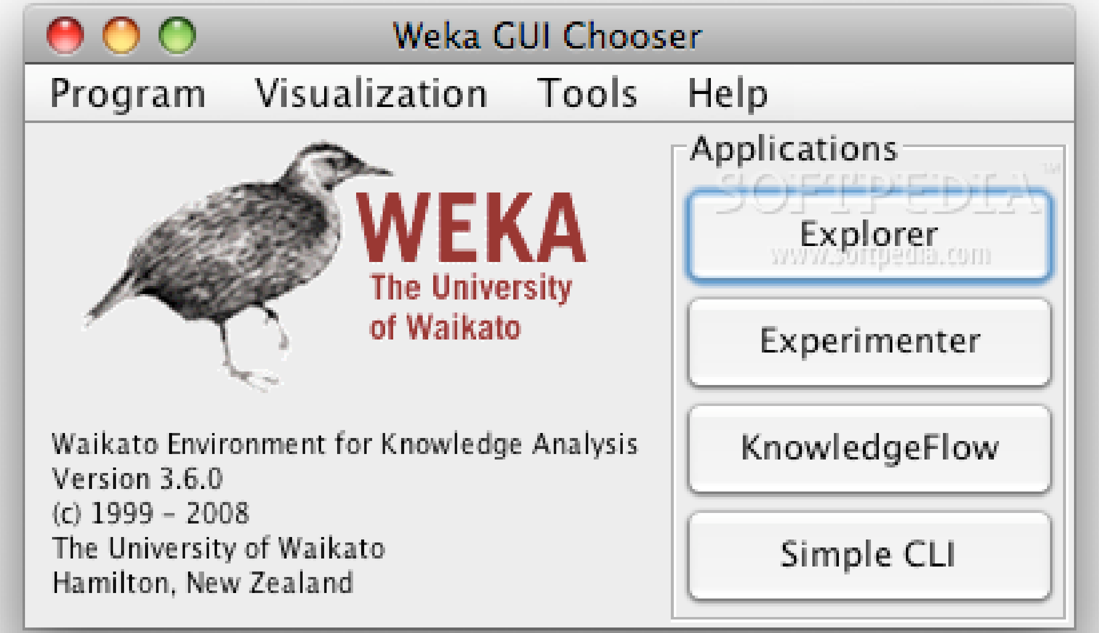
Classifier output

```
Time taken to build model: 0.24 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0         1         1         1         Iris-setosa
  1         0.063     0.905     1         0.95      Iris-versicolor
  0.882     0         1         0.882     0.938     Iris-virginica

=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  0  0 |  a = Iris-setosa
  0 19  0 |  b = Iris-versicolor
  0  2 15 |  c = Iris-virginica
```

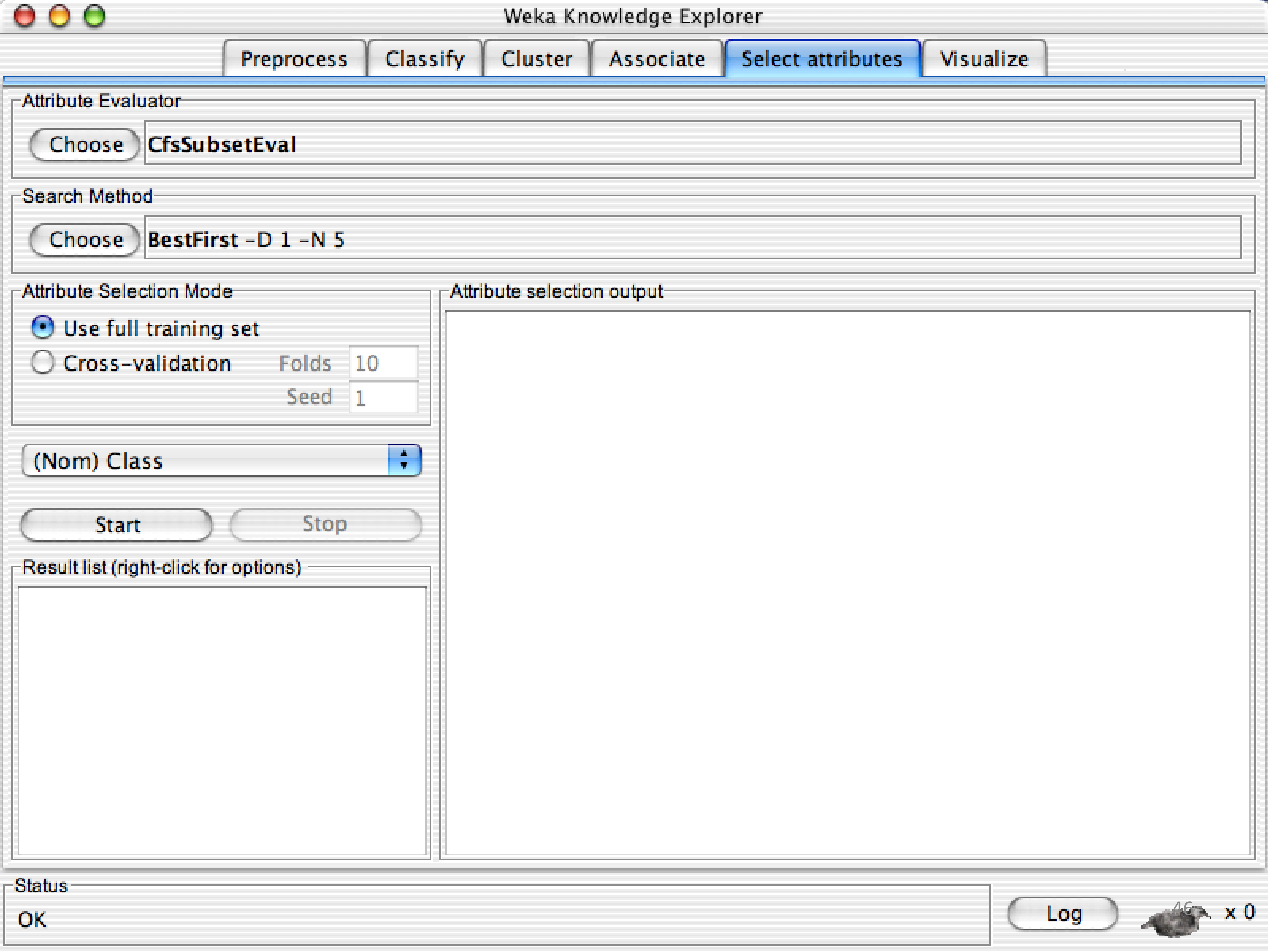
Status

OK

Weka Attribute Selection Steps



Waikato Environment for Knowledge Analysis
Version 3.6.0
(c) 1999 - 2008
The University of Waikato
Hamilton, New Zealand



Attribute Evaluator

CfsSubsetEval

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

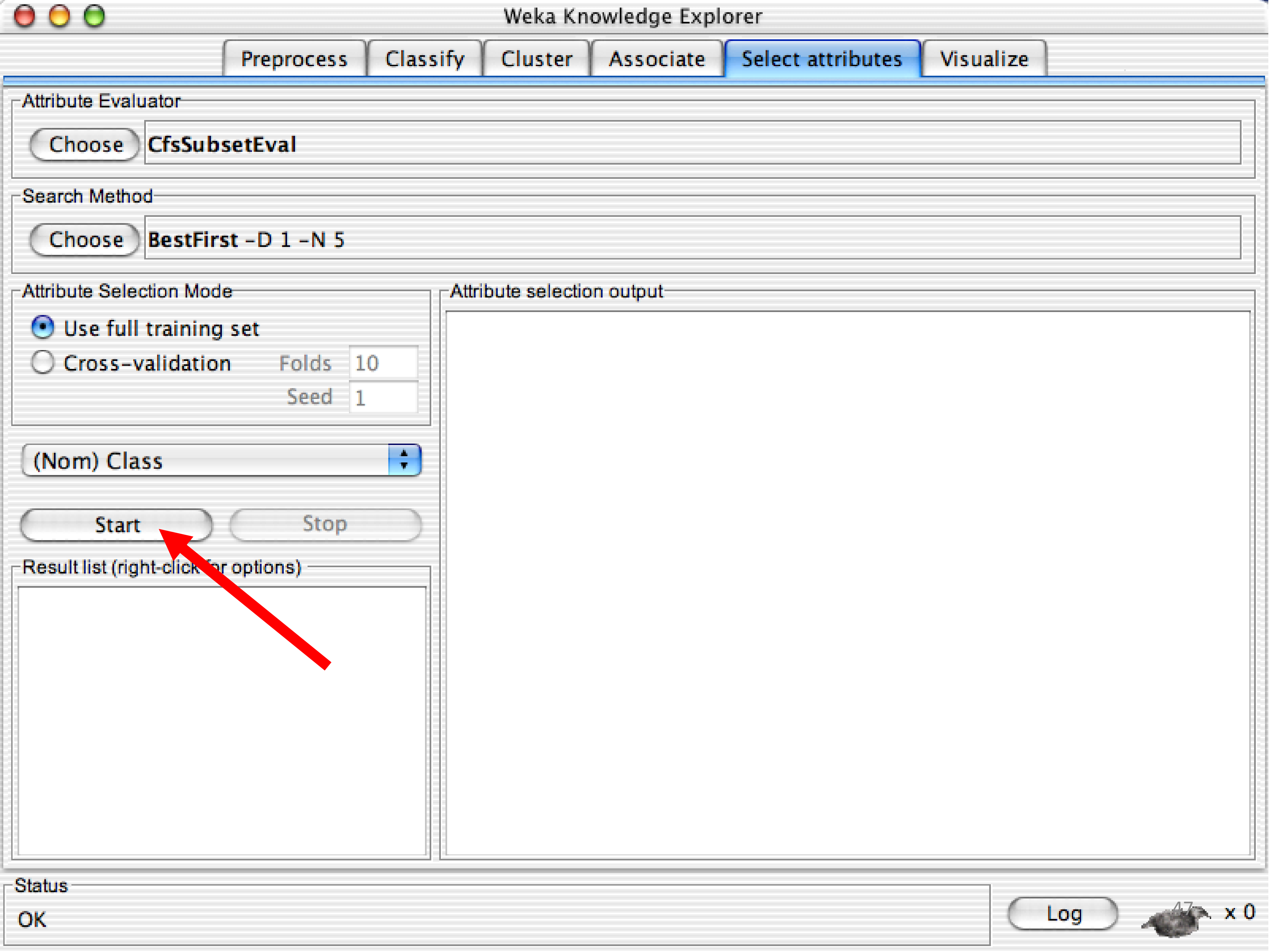
○ Cross-validation Folds 10

Seed 1

(Nom) Class

Status

OK



Attribute Evaluator: CfsSubsetEval

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode

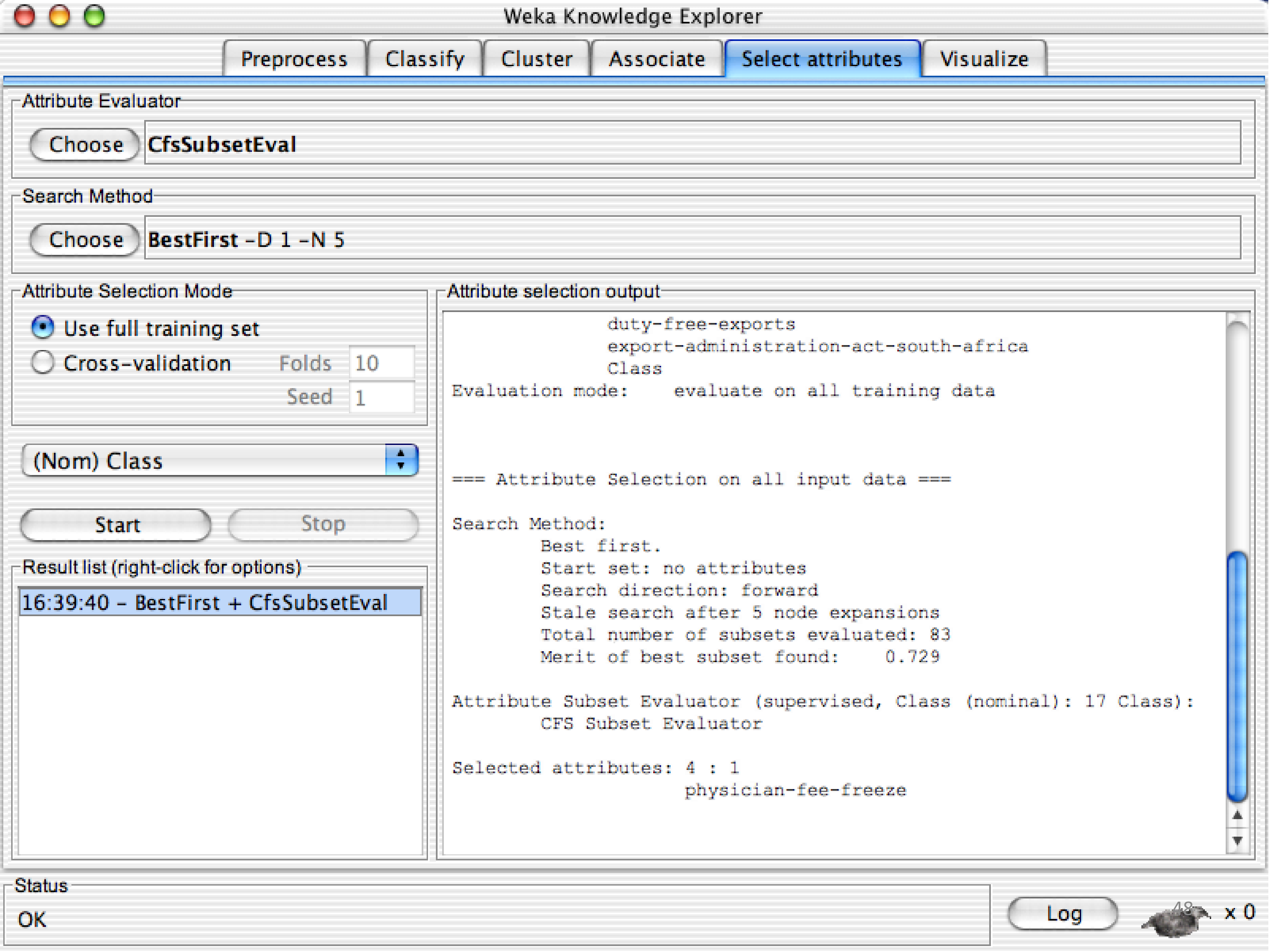
● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

Status

OK



Attribute Evaluator

CfsSubsetEval

Search Method

BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) Class

Result list (right-click for options)

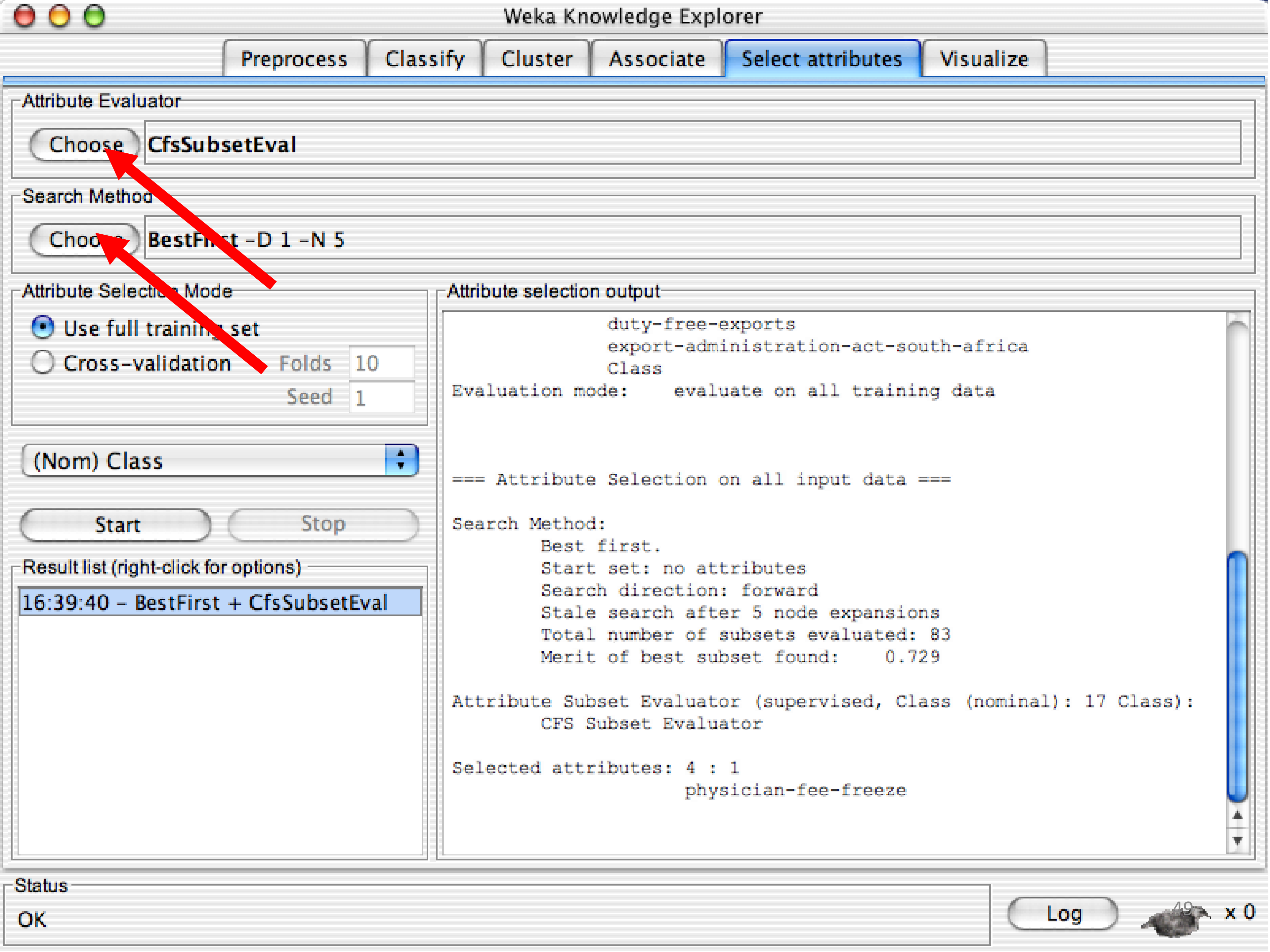
16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

```
                duty-free-exports
                export-administration-act-south-africa
                Class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 83
        Merit of best subset found:    0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
        CFS Subset Evaluator

Selected attributes: 4 : 1
                     physician-fee-freeze
```

Status

OK



Attribute Evaluator: CfsSubsetEval

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

Result list (right-click for options)

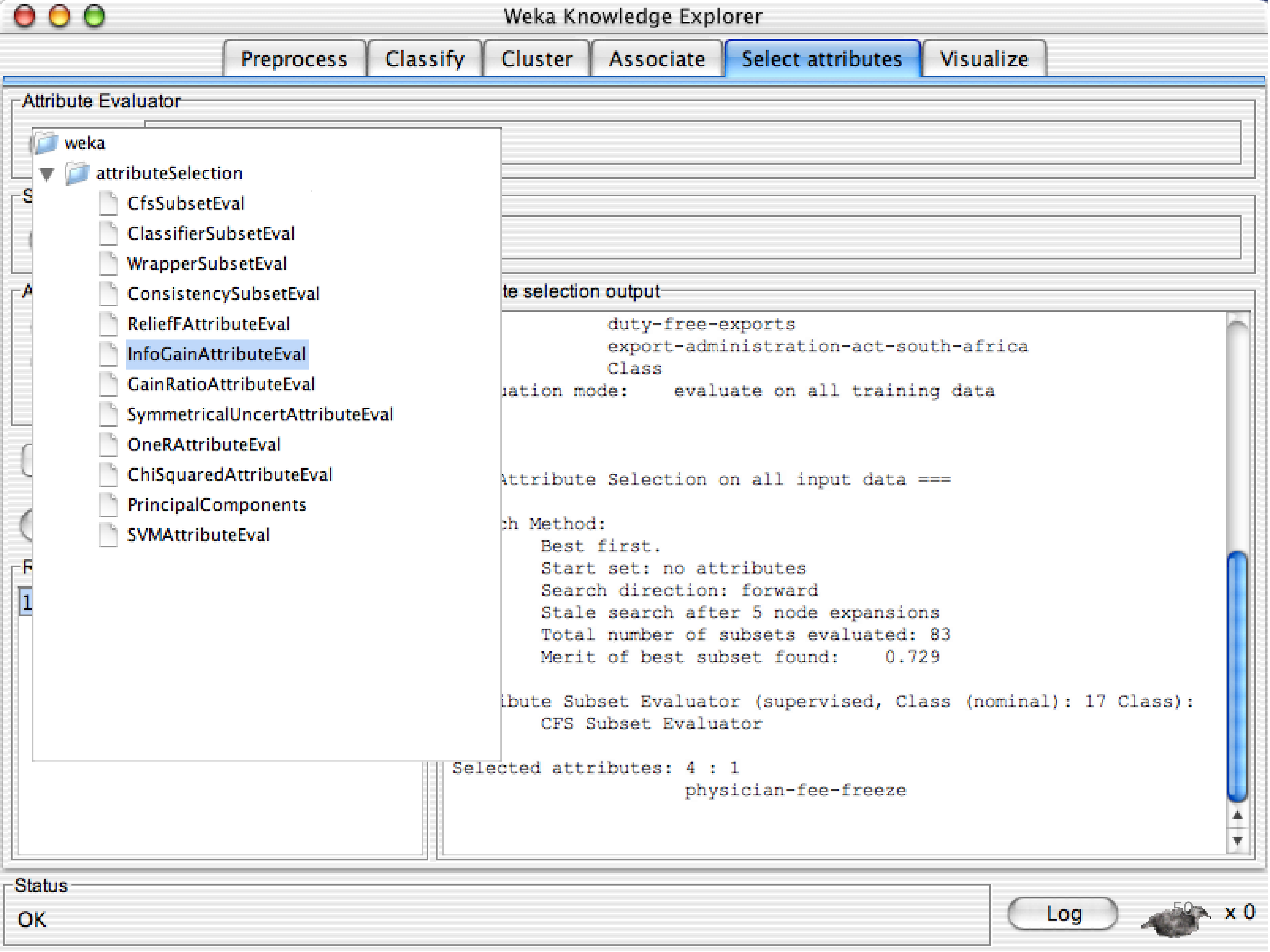
16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

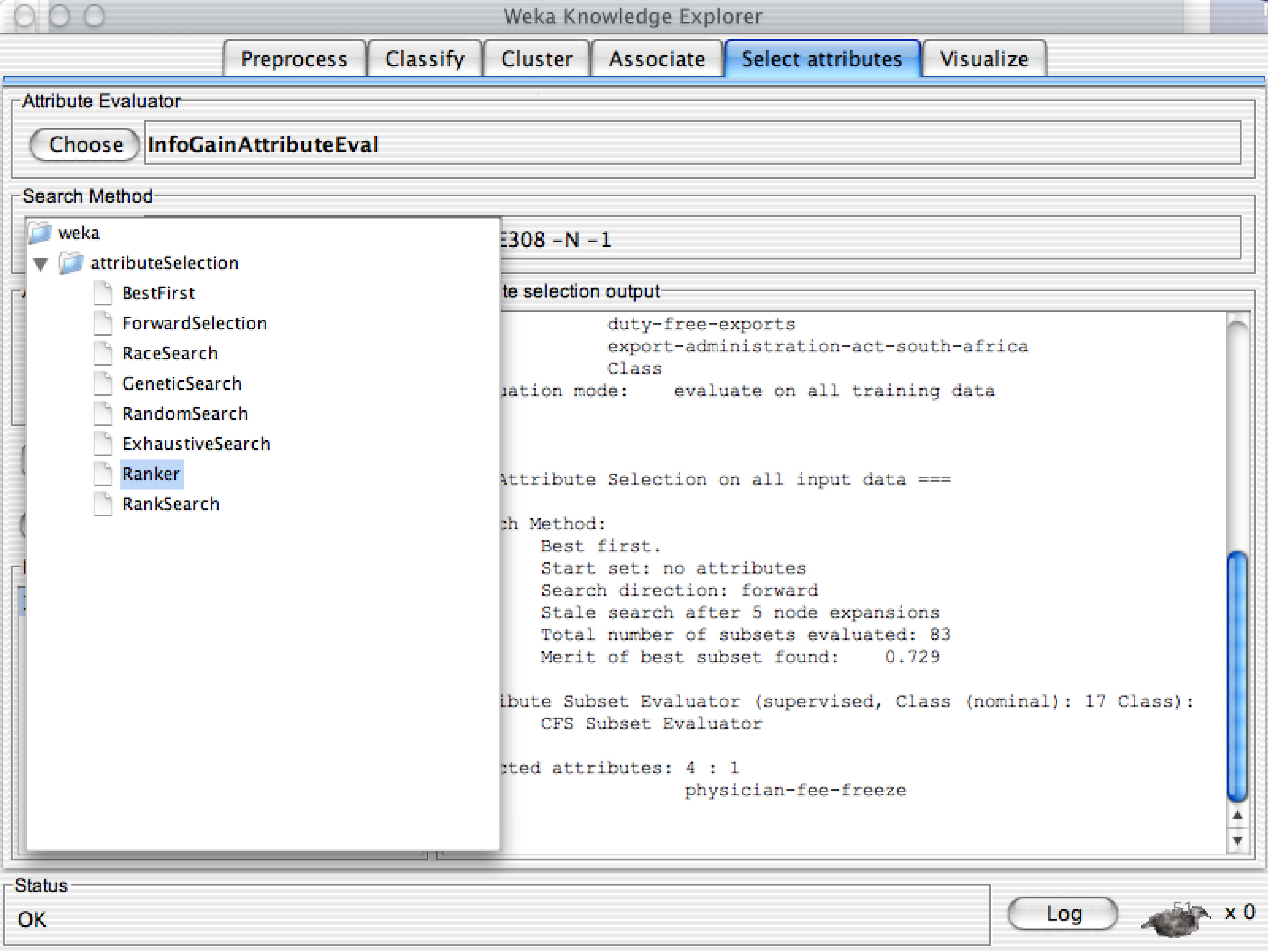
```
                duty-free-exports
                export-administration-act-south-africa
                Class
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 83
        Merit of best subset found:    0.729

Attribute Subset Evaluator (supervised, Class (nominal): 17 Class):
        CFS Subset Evaluator

Selected attributes: 4 : 1
                     physician-fee-freeze
```

Status

OK



```
             duty-free-exports
             export-administration-act-south-africa
             Class
uation mode:    evaluate on all training data



Attribute Selection on all input data ===

ch Method:
       Best first.
       Start set: no attributes
       Search direction: forward
       Stale search after 5 node expansions
       Total number of subsets evaluated: 83
       Merit of best subset found:    0.729

ibute Subset Evaluator (supervised, Class (nominal): 17 Class):
       CFS Subset Evaluator

Selected attributes: 4 : 1
                     physician-fee-freeze
```



Attribute Evaluator

InfoGainAttributeEval

Search Method

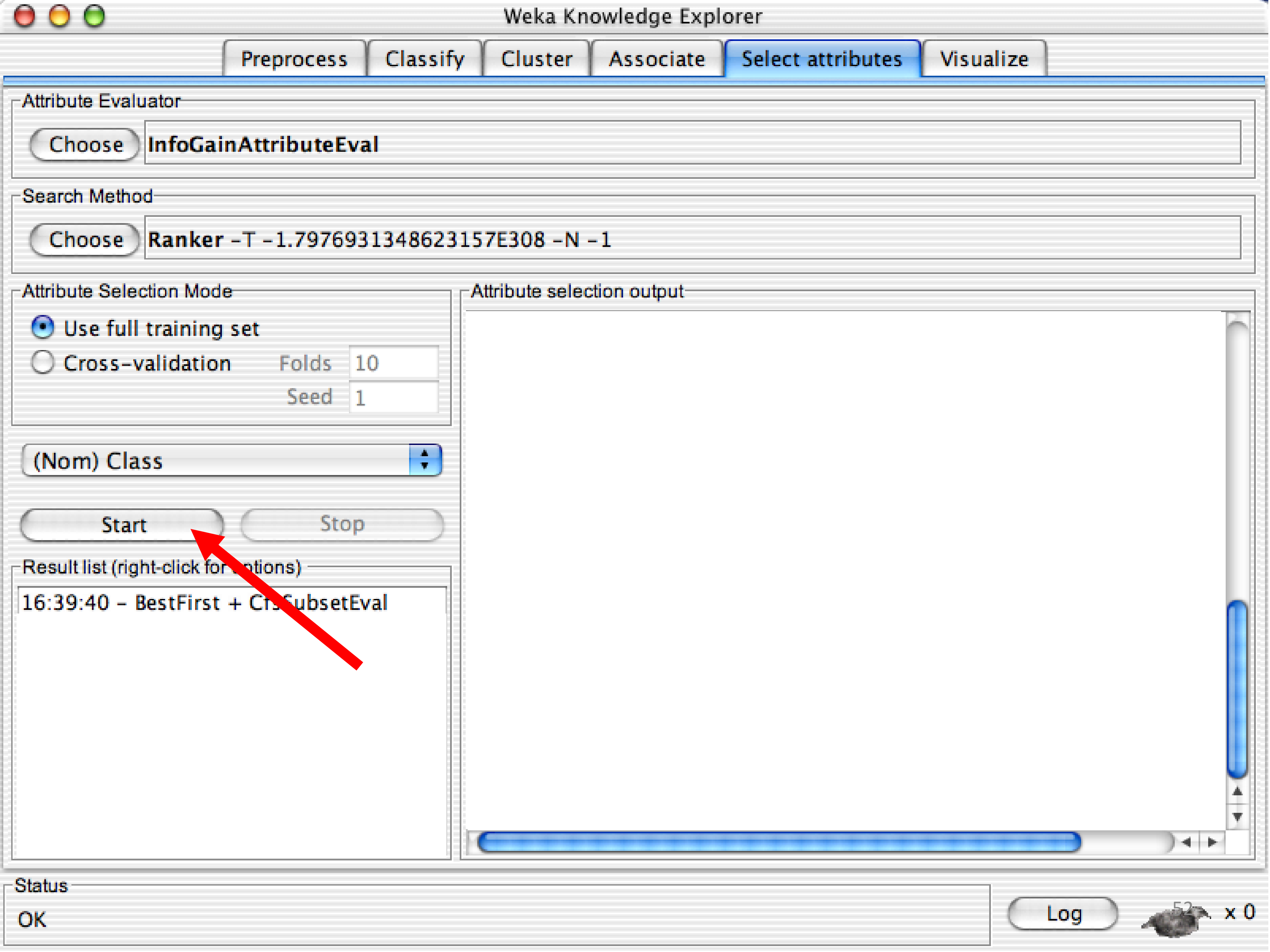
E308 -N -1

te selection output

```
              duty-free-exports
              export-administration-act-south-africa
              Class
uation mode:    evaluate on all training data



Attribute Selection on all input data ===

ch Method:
     Best first.
     Start set: no attributes
     Search direction: forward
     Stale search after 5 node expansions
     Total number of subsets evaluated: 83
     Merit of best subset found:    0.729

ibute Subset Evaluator (supervised, Class (nominal): 17 Class):
     CFS Subset Evaluator

cted attributes: 4 : 1
                  physician-fee-freeze
```

Status

OK



Attribute Evaluator

InfoGainAttributeEval

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10

Seed 1

(Nom) Class

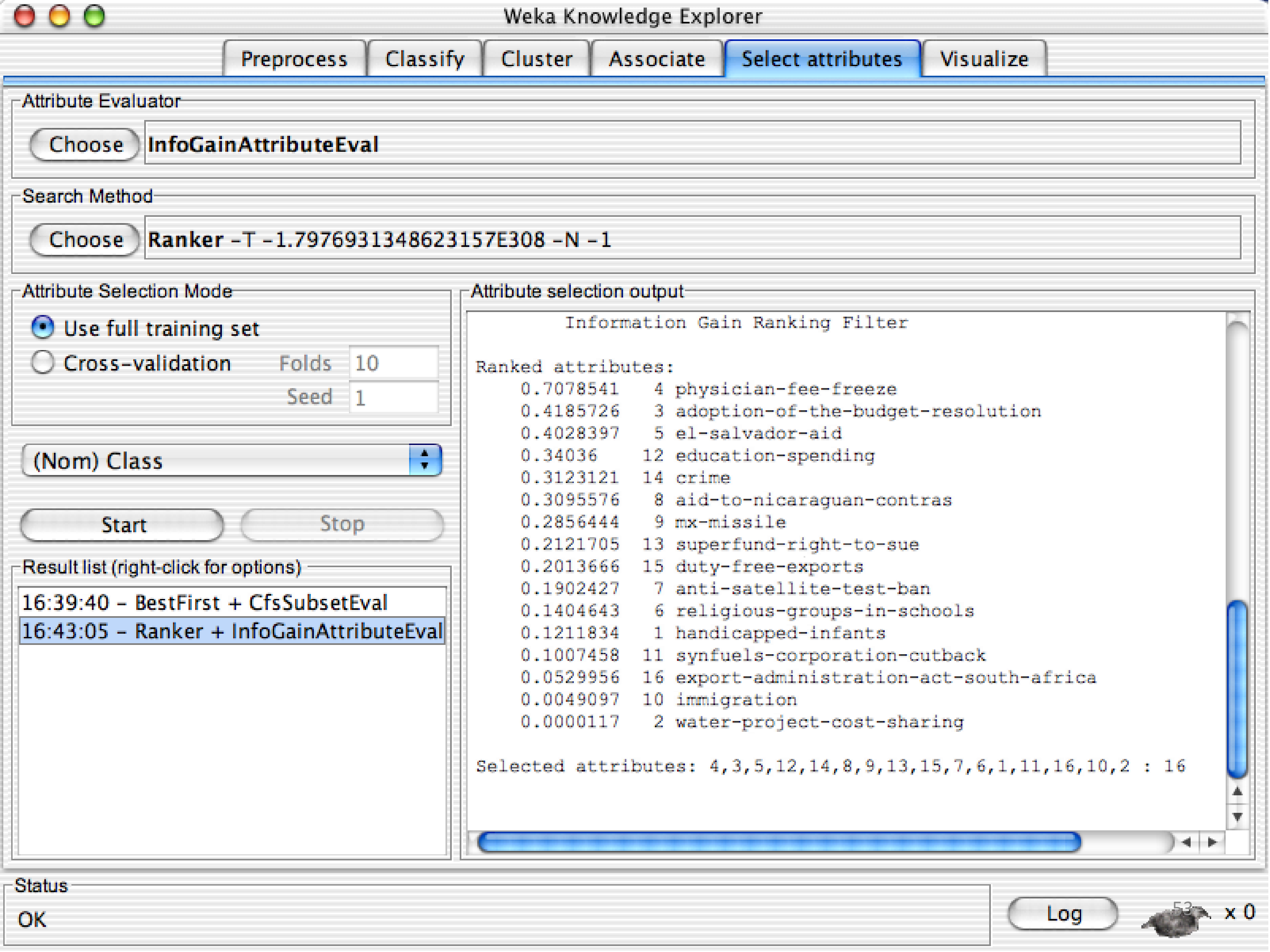
Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

Attribute selection output

Status

OK



Attribute Evaluator

InfoGainAttributeEval

Search Method

Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode

● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) Class

Result list (right-click for options)

16:39:40 - BestFirst + CfsSubsetEval

16:43:05 - Ranker + InfoGainAttributeEval

Attribute selection output

```
        Information Gain Ranking Filter

Ranked attributes:
    0.7078541   4 physician-fee-freeze
    0.4185726   3 adoption-of-the-budget-resolution
    0.4028397   5 el-salvador-aid
    0.34036    12 education-spending
    0.3123121  14 crime
    0.3095576   8 aid-to-nicaraguan-contras
    0.2856444   9 mx-missile
    0.2121705  13 superfund-right-to-sue
    0.2013666  15 duty-free-exports
    0.1902427   7 anti-satellite-test-ban
    0.1404643   6 religious-groups-in-schools
    0.1211834   1 handicapped-infants
    0.1007458  11 synfuels-corporation-cutback
    0.0529956  16 export-administration-act-south-africa
    0.0049097  10 immigration
    0.0000117   2 water-project-cost-sharing

Selected attributes: 4,3,5,12,14,8,9,13,15,7,6,1,11,16,10,2 : 16
```

Status

OK

Thank You