Data Mining III – Lesson 5

Tamara B. Sipes, Ph.D.

Last Time: Lesson 4

- Continued mining the Image Segmentation dataset
- Transitioned to more complex data mining tasks
- Introduction to several of the common issues when mining real-world datasets, such as having a balanced dataset, sampling, expanding, etc.
- Focused on model building for best novel data/ unseen data evaluation results

Lesson 4: The Lesson Learned

- The importance of having a balanced data
- How to create a balanced dataset
- The use of your data mining expertise earlier in the process (we could have abandoned all the black-box models earlier)
- Got introduced to the notion of “cost” and “cost effective” data mining
- Your knowledge about the methods, how they work, what they can accomplish, what kind of data they are best suited for, is essential in the DM process
- We built up your experience and intuition

Lesson 5 Overview

- Continuing with more complex data mining tasks
- Introduction to several of the commonly encountered difficulties when mining real-world datasets
- Focus on modeling difficult datasets
- Exposure to “dirty” data and all the issues that arise from it
- Get more practical experience

Instructions

- Hands-on lesson #4
- Please be ready to have both the Lesson 5 opened and the weka Explorer as well
- Follow the step by step instructions, and perform the modeling of the dataset yourself as well, to gain more hands-on data mining experience

- Let's get started!

Dataset

- Open Shoe.arff
- TO DO:
  - Open weka 3.5.7 or similar
  - Choose Explorer from the Applications menu
  - “Open file” Shoe.arff

Data and Problem Description

- A national shoe chain wants to model customer profiles in order to better understand their market
- More than 26,600 customer-purchase profiles are collected from their national chain of shoe stores
- 14 attributes/variables/descriptors
- #instances: 26608

Variables

- Gender
- City
- State
- ZIP
- StoreCd
- Source
- Age
- YearsRunning
- Miles/Week
- Races/Year
- Triathlete
- Style
- ShoeCode
- PurchaseNumber

Problem Description

- The goal:

  "To understand customer motivations for purchasing shoes"

- Is the collected information able to help the shoe company understand customer motivations?

Classification Task

- Model the "PurchaseNumber" – the number of shoes a customer has purchased
- Slight problem: no records of the customers that have not purchased any shoes, only those that have purchased
- Population not perfectly sampled/represented
- A quick examination of the variable shows only two distinct values: 1 and 2

Let's Get Started

- The next step after the problem definition is data preparation
- First, we need to examine all the variables, input and output
- 13 input variables
- 1 output variable
- Side note: is it possible to model a dataset with more than one output variable? Yes.

Looking at the Variables: Gender

Selected attribute

Name: Gender
Missing: 0 (0%)
Distinct: 6
Type: Nominal
Unique: 1 (0%)

Label	Count
M	13348
F	10361
U	2420
?	445
A	33
7.0	1

Gender: Analysis

- Six distinct values for gender:

  M, F, U, ?, A, 7.0

- Indicates a problem, there should be at most three (M, F, missing)
- To do:
  - Merge U and ?
  - Examine “A” – “absent” or something else?
  - Examine “7.0” – look at the whole record

Gender: Action Items

- Merge “U” and “?”:

  weka.filters.unsupervised.attribute.MergeTwoValues -C first -F 3 -S 4

Selected attribute

Name: Gender Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 1 (0%)

Label	Count
M	13348
F	10361
U_?	2865
A	33
7.0	1

Gender: Action Items

- The examination of the “A” records showed mostly empty attribute values
- Must be some label of the “absent” data items
- Remove all instances with “A” in Gender:

  weka.filters.unsupervised.instance.RemoveWithValues -S 0.0 -C first -L 4 -H

Gender: Result

Selected attribute

Name: Gender
Missing: 0 (0%)
Distinct: 4
Type: Nominal
Unique: 1 (0%)

Label	Count
M	13348
F	10361
U_?	2865
7.0	1

Gender: More To Do

- Examing “7.0” – OK
- Merge it with the rest of the missing:

  weka.filters.unsupervised.attribute.MergeTwoValues -C first -F 3 -S 4
- “U_?_7.0” is really a missing value, not another nominal gategory
- Rename the “U_?_7.0” so that it is “’?’”:
- First, save the file as Shoe1.arff

Renaming a Nominal Value

- Rename the “U_?_7.0” so that it is “’?’”:
- Open Shoe1.arff in Wordpad
- In: @attribute Gender {M,F,U_?_7.0}, change U_?_7.0 to ‘?’ (with the quotes), and all the occurrences
- If you enter just ?, you will get an error – nominal missing values need to be labeled with ‘?’ in weka
- Save as Shoe2.arff (why?)
- Reopen Shoe2.arff in weka

Gender: Completed

- Let's examine Gender now:

Selected attribute

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

Label	Count
M	13348
F	10361
?	2866

Looking at the Variables: City

Selected attribute

Name: City
Missing: 0 (0%)
Distinct: 4753
Type: Nominal
Unique: 2252 (8%)

Label	Count
?	1
ANN ANBOR	2
AGAWAM	1
AMHERST	14
BELCHERTOWN	4
CHICOPEE	10
EASTHAMPTON	7
SOUTHWICK	2
E LONGMEADOW	4
EAST LONGMEADOW	1
EAST LONG	1
FEEDING HILLS	3

City: Analysis

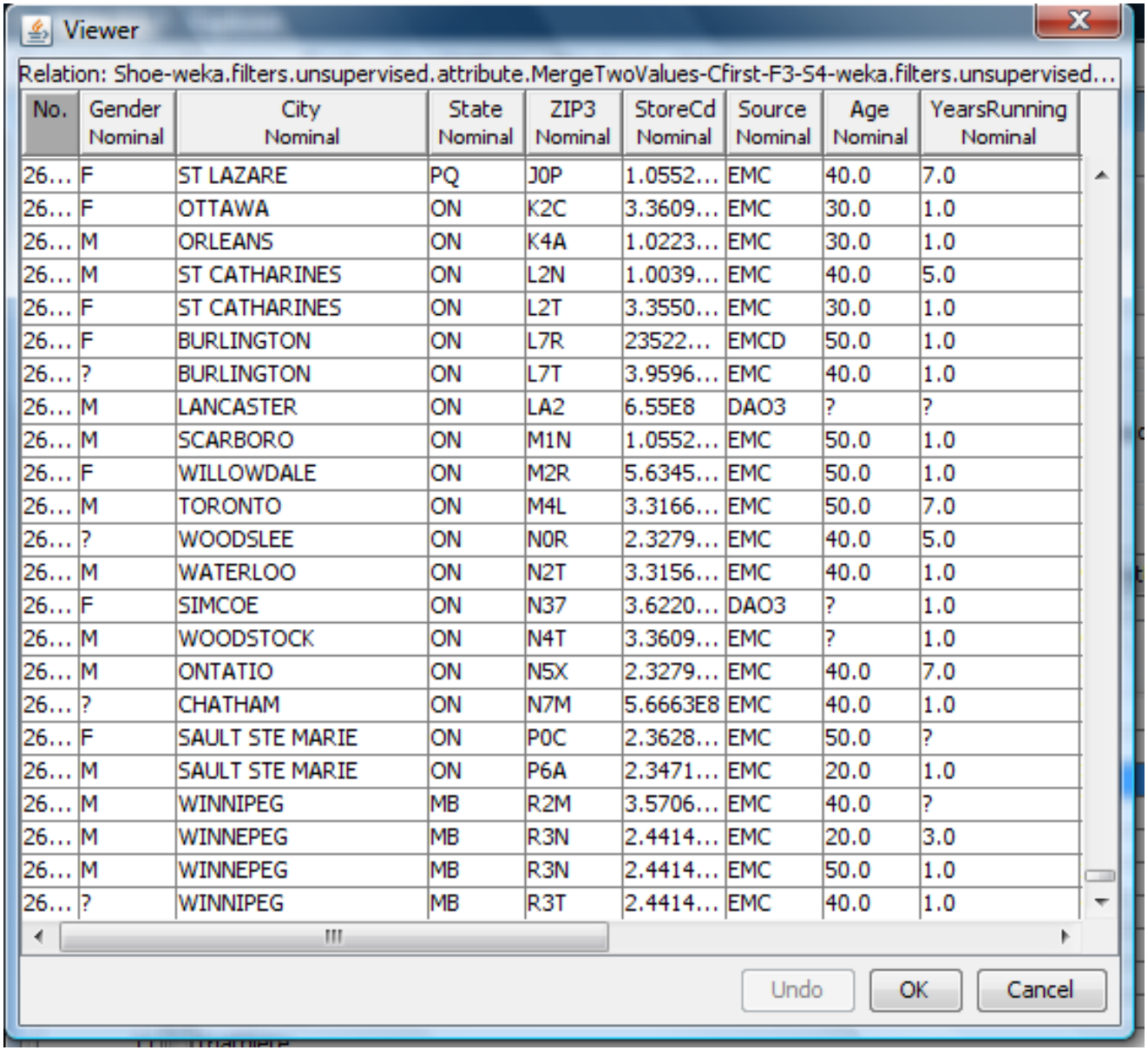
- Lots of problems
- 4753 distinct values, 2252 unique
- Not very predictive
- Obvious typos: E Longmeadow, East Longmeadow, East Long, for example, and many others
- There is also the zip code variable, which could be enough to represent the geographical area

City: Action Items

- It would be too time consuming to examine/change all the typos
- Remove City?
- Let's wait and see, we will look at the rest of the data first

Examine State

- 61 states?
- Let's examine the data



No.	Gender Nominal	City Nominal	State Nominal	ZIP3 Nominal	StoreCd Nominal	Source Nominal	Age Nominal	YearsRunning Nominal
26...	F	ST LAZARE	PQ	J0P	1.0552...	EMC	40.0	7.0
26...	F	OTTAWA	ON	K2C	3.3609...	EMC	30.0	1.0
26...	M	ORLEANS	ON	K4A	1.0223...	EMC	30.0	1.0
26...	M	ST CATHARINES	ON	L2N	1.0039...	EMC	40.0	5.0
26...	F	ST CATHARINES	ON	L2T	3.3550...	EMC	30.0	1.0
26...	F	BURLINGTON	ON	L7R	23522...	EMCD	50.0	1.0
26...	?	BURLINGTON	ON	L7T	3.9596...	EMC	40.0	1.0
26...	M	LANCASTER	ON	LA2	6.55E8	DAO3	?	?
26...	M	SCARBORO	ON	M1N	1.0552...	EMC	50.0	1.0
26...	F	WILLOWDALE	ON	M2R	5.6345...	EMC	50.0	1.0
26...	M	TORONTO	ON	M4L	3.3166...	EMC	50.0	7.0
26...	?	WOODSLEE	ON	N0R	2.3279...	EMC	40.0	5.0
26...	M	WATERLOO	ON	N2T	3.3156...	EMC	40.0	1.0
26...	F	SIMCOE	ON	N37	3.6220...	DAO3	?	1.0
26...	M	WOODSTOCK	ON	N4T	3.3609...	EMC	?	1.0
26...	M	ONTATIO	ON	N5X	2.3279...	EMC	40.0	7.0
26...	?	CHATHAM	ON	N7M	5.6663E8	EMC	40.0	1.0
26...	F	SAULT STE MARIE	ON	P0C	2.3628...	EMC	50.0	?
26...	M	SAULT STE MARIE	ON	P6A	2.3471...	EMC	20.0	1.0
26...	M	WINNIPEG	MB	R2M	3.5706...	EMC	40.0	?
26...	M	WINNEPEG	MB	R3N	2.4414...	EMC	20.0	3.0
26...	M	WINNEPEG	MB	R3N	2.4414...	EMC	50.0	1.0
26...	?	WINNIPEG	MB	R3T	2.4414...	EMC	40.0	1.0

State: Analysis

- Visually examine the data, either using Edit or in Excel
- The unusual states were the out of town buyers (Canada etc.)
- There are also some obvious foreign visitors with unusual states and zip codes
- There are some unusual zip code values associated with the out of country state values

ZIP3: Analysis

- Three-digit zip codes
- First three digits, the store chain said
- Careful examination of the data revealed that there are some '999'
- The 999 values represent missing zip codes (or, ?) – there are 81 entries with the zip3 = 999
- The alphanumeric zip3 values correspond to the Canadian zip codes
- F10, F11, F16, F20 and F29 correspond to the international "zip codes"

Zip3: Action Items

- Replace 999.0 with ? Edit/Replace values with
- Make sure that no other 999 values are replaced!
- Press OK to exit the Edit window
- Save as shoe3.arff
- Open shoe3.arff in weka
- Check the count of ?: it was 3, it is 84 now

Zip3: Continue the Analysis

- Now 831 distinct values
- City has 4753!
- One city could have more than one zip code associated with it, what about the first three digits (zip3)?
- Let's delete city:
- Select City, click Remove
- Save the current dataset

StoreCd: Analysis

Selected attribute

Name: StoreCd Type: Nominal
Missing: 0 (0%) Distinct: 1211 Unique: 280 (1%)

Label	Count
4.309416E9	116
2366530.0	7
2.22624E9	83
2.271016E9	45
2230950.0	6
2.23095E9	35
2226240.0	12
7.55029E8	820
4.657555E9	25
2.223665E9	29
?	3716
2271016.0	2

StoreCd: Results

- 1211 distinct values
- 3716 '?' values
- Values all over the place
- Good to have a geographic information represented, but going to the store granularity is most likely too much for this problem
- Then again, it could be that certain stores promote more shoe sales!
- Numeric values – messed up, remove the variable

Source: Analysis

- 25 distinct values
- Seems like there is some redundancy, like in DAOA and DAOa, DAOC, DAO4?
- But, there is DAO3 too?
- More investigation needed
- Need to consult the domain experts

Triathlete: Examine

Selected attribute

Name: Triathlete
Missing: 0 (0%)
Distinct: 5
Type: Nominal
Unique: 1 (0%)

Label	Count
?	5197
N	19119
Y	2256
6.0	1
0.0	2

Triathlete: Analysis

- Values 6.0 and 0.0 incorrect
- You can actually remove them, or change to ?
- weka.filters.unsupervised.instance.RemoveWithValues -S 0.0 -C 9 -L 4-5 –H

ShoeCode

- These change all the time
- Too many distinct values as well
- Remove

Save

- Save the file as Shoe4.arff

PurchaseNumber: Examine

Selected attribute

Name: PurchaseNumber Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	2
Mean	1.057
StdDev	0.233

Nominal Prediction

- weka.filters.unsupervised.attribute.NumericToNominal -R last
- Save as Shoe5.arff

PurchaseNumber

Selected attribute

Name: PurchaseNumber Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Label	Count
1	25047
2	1525

Balanced Data

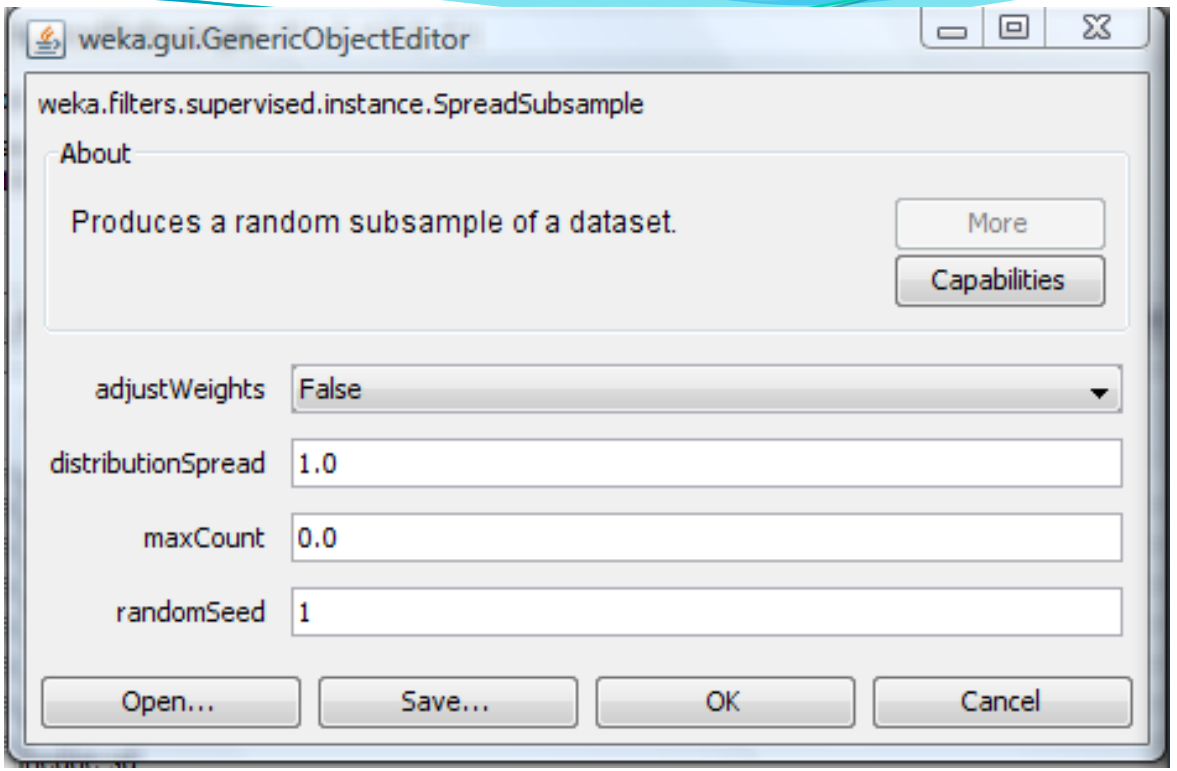
- Having a balanced dataset means that there is approximately the same number of each types of examples represented to the learning method
- In other words, the distribution of the class variable values should be even
- Shoe dataset – not balanced data

How to Balance a Dataset?

- Two options:
  - Subsample
  - Expand

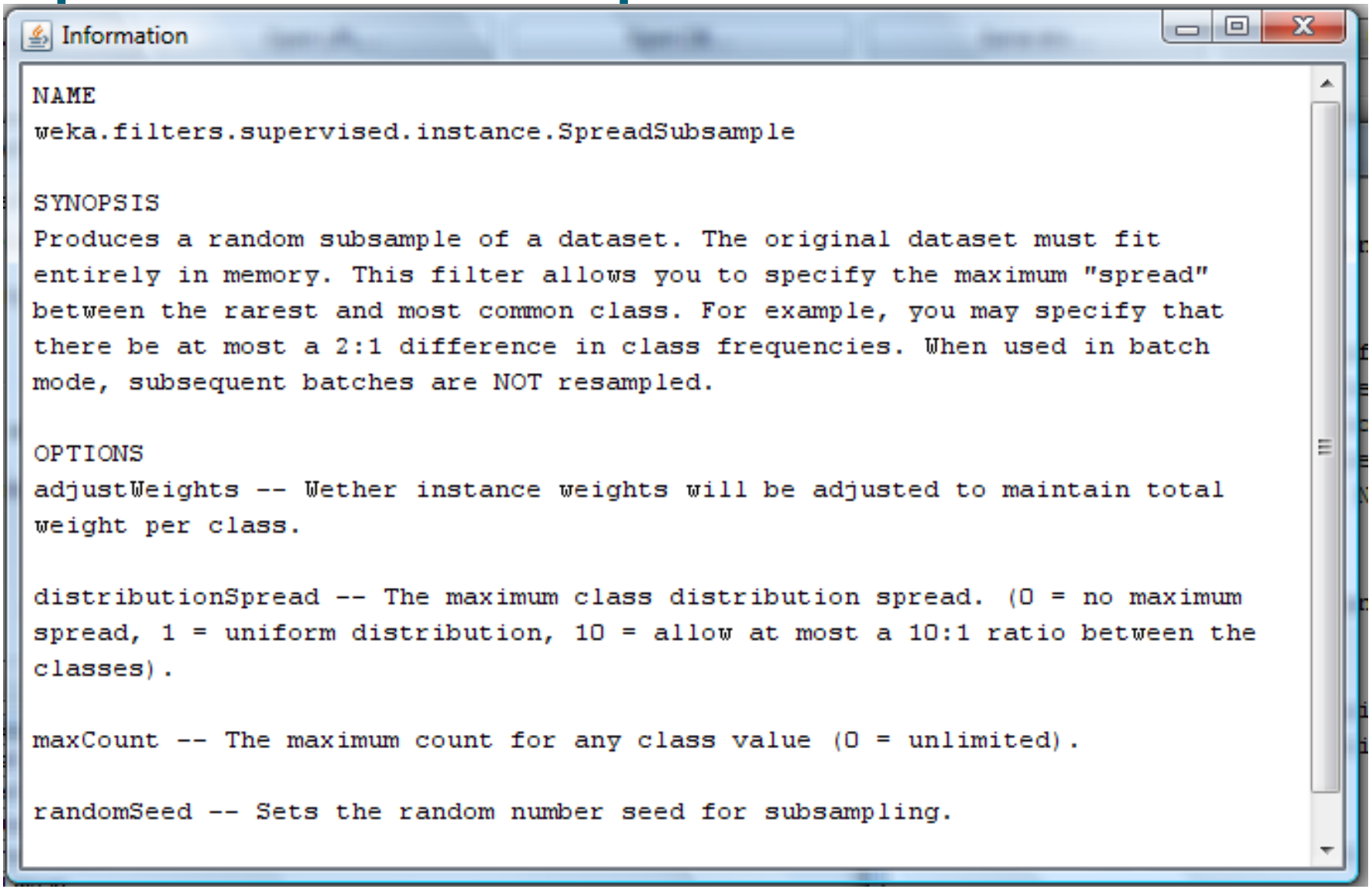
- What to do?
- Too much “1” data – let’s subsample

Sampling



- Filter: SpreadSubsample

SpreadSubsample: Parameters



SpreadSubsample: The Call

- weka.filters.supervised.instance.SpreadSubsample -M 2.0 -X 0.0 -S 1

  (distributionSpread = 2:1 at most)

- Why 2:1?
- Click “Apply”
- Number of instances: 4574

PurchaseNumber:

Selected attribute

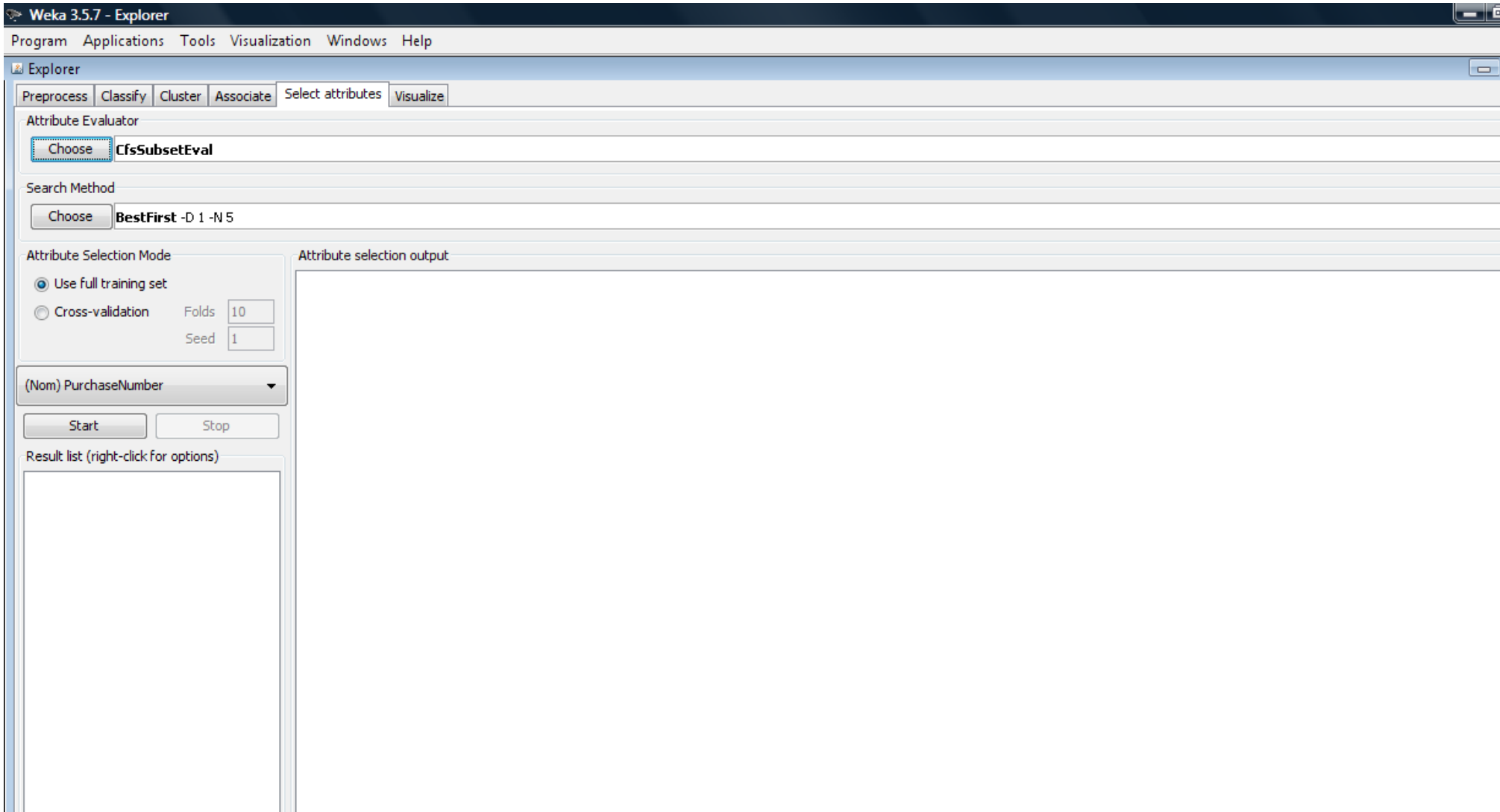
Name: PurchaseNumber
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

Label	Count
1	3050
2	1525

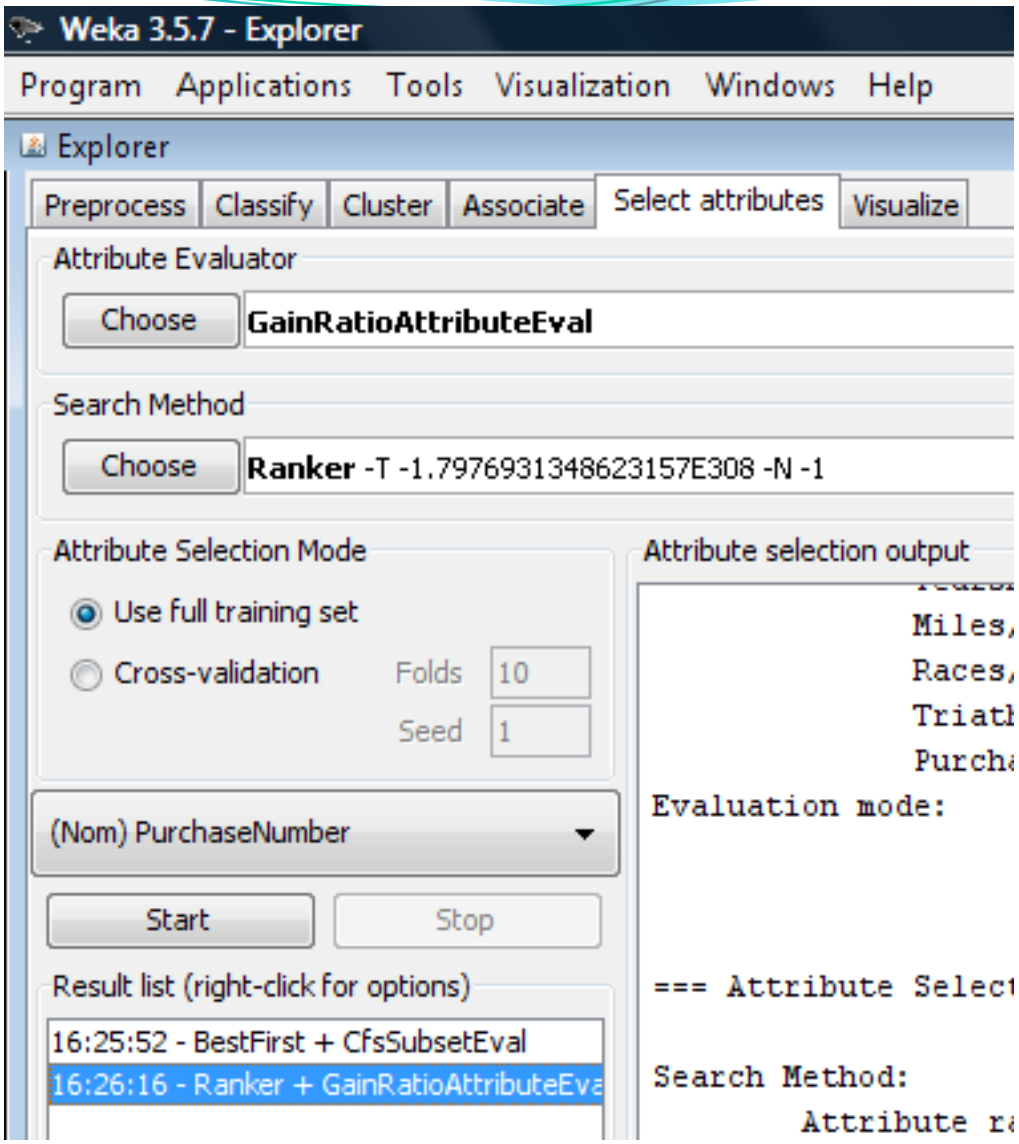
What's Next?

- Let's evaluate the attributes
- The dataset is not perfectly clean
- Let's get some help

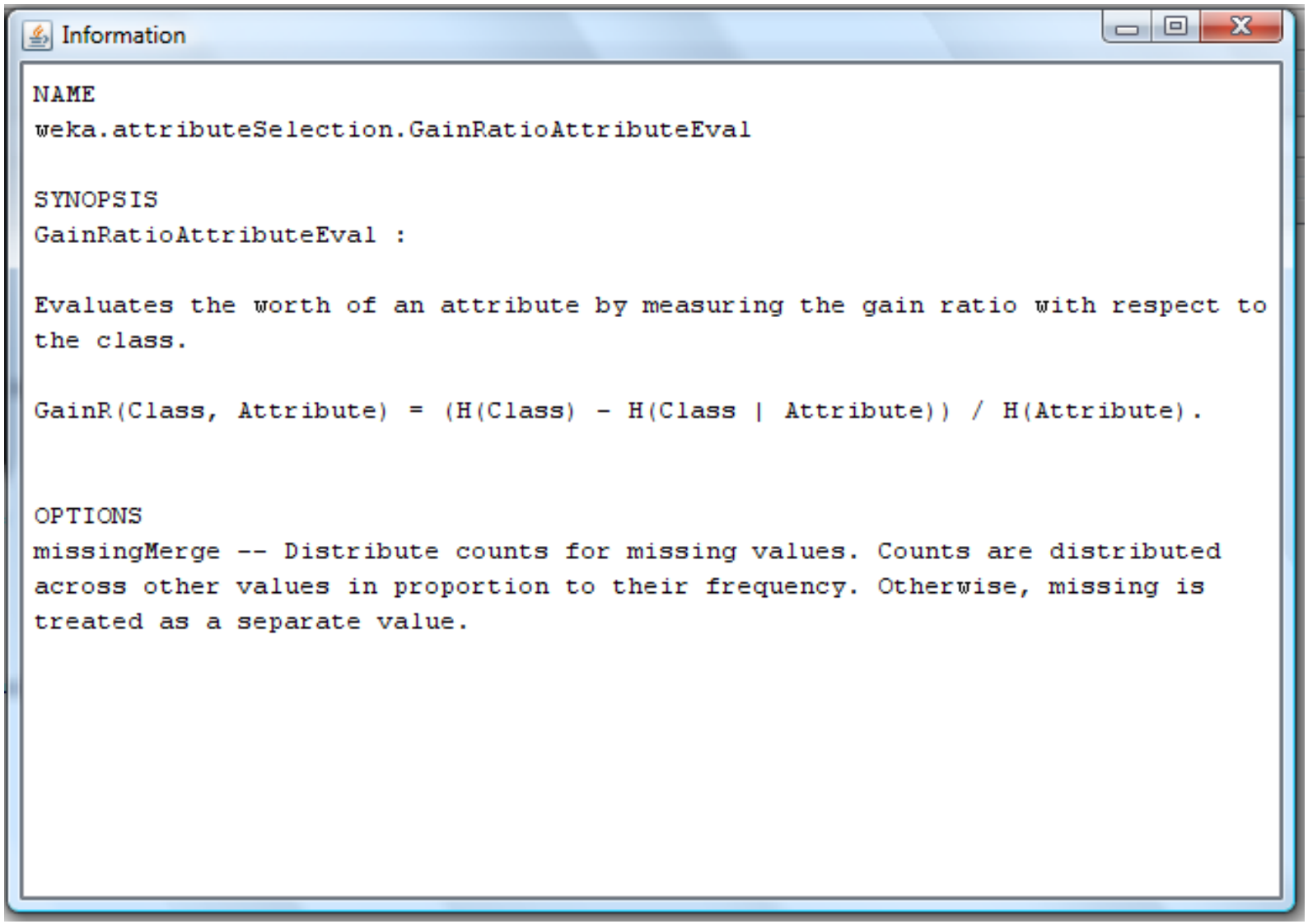
Attribute Selection Tab



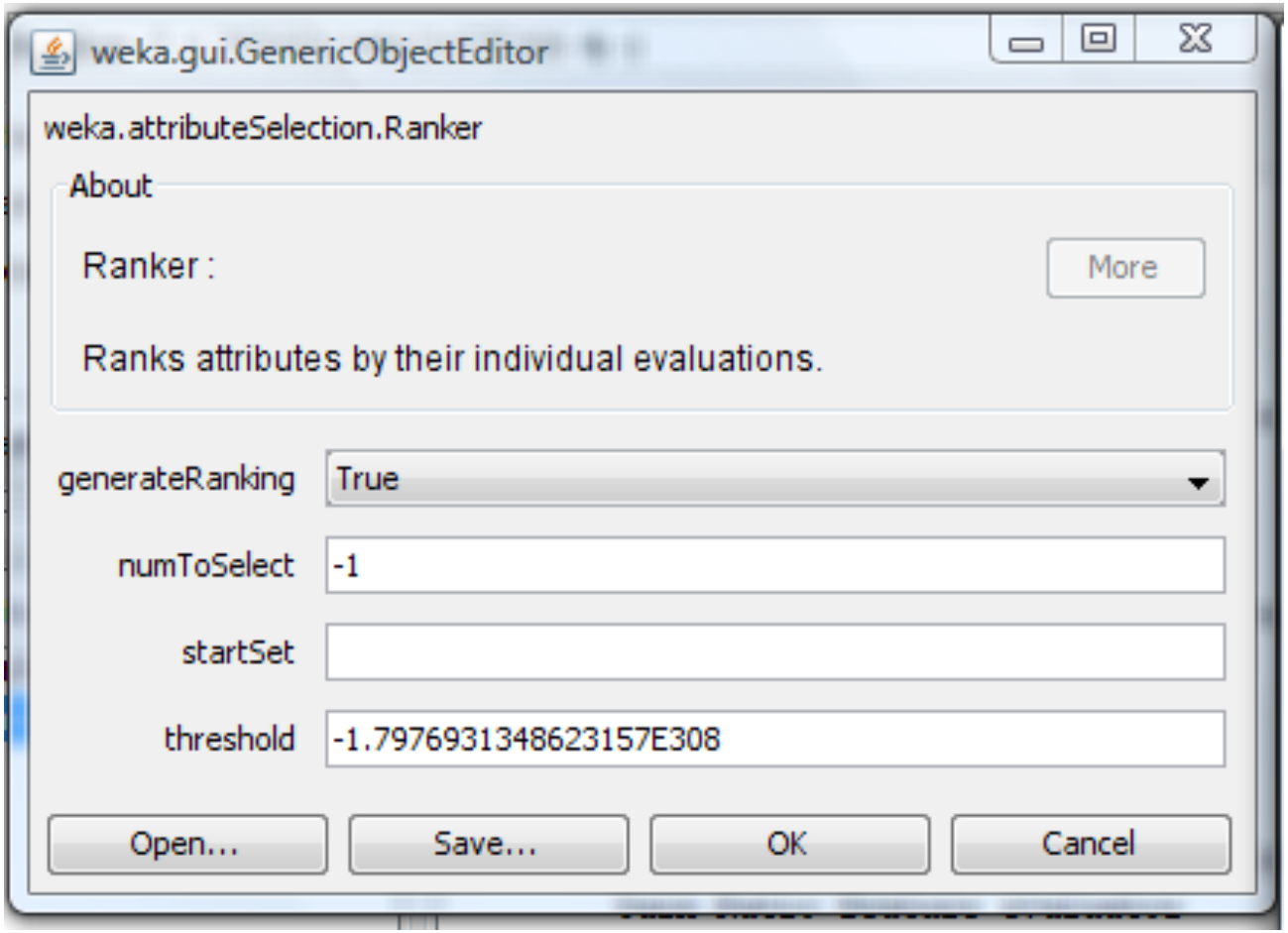
GainRatio


Weka 3.5.7 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Attribute Evaluator
Choose GainRatioAttributeEval
Search Method
Choose Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation Folds 10
Seed 1
(Nom) PurchaseNumber
Start Stop
Result list (right-click for options)
16:25:52 - BestFirst + CfsSubsetEval
16:26:16 - Ranker + GainRatioAttributeEva
Attribute selection output
Miles,
Races,
Triath
Purcha
Evaluation mode:
=== Attribute Select
Search Method:
Attribute ra

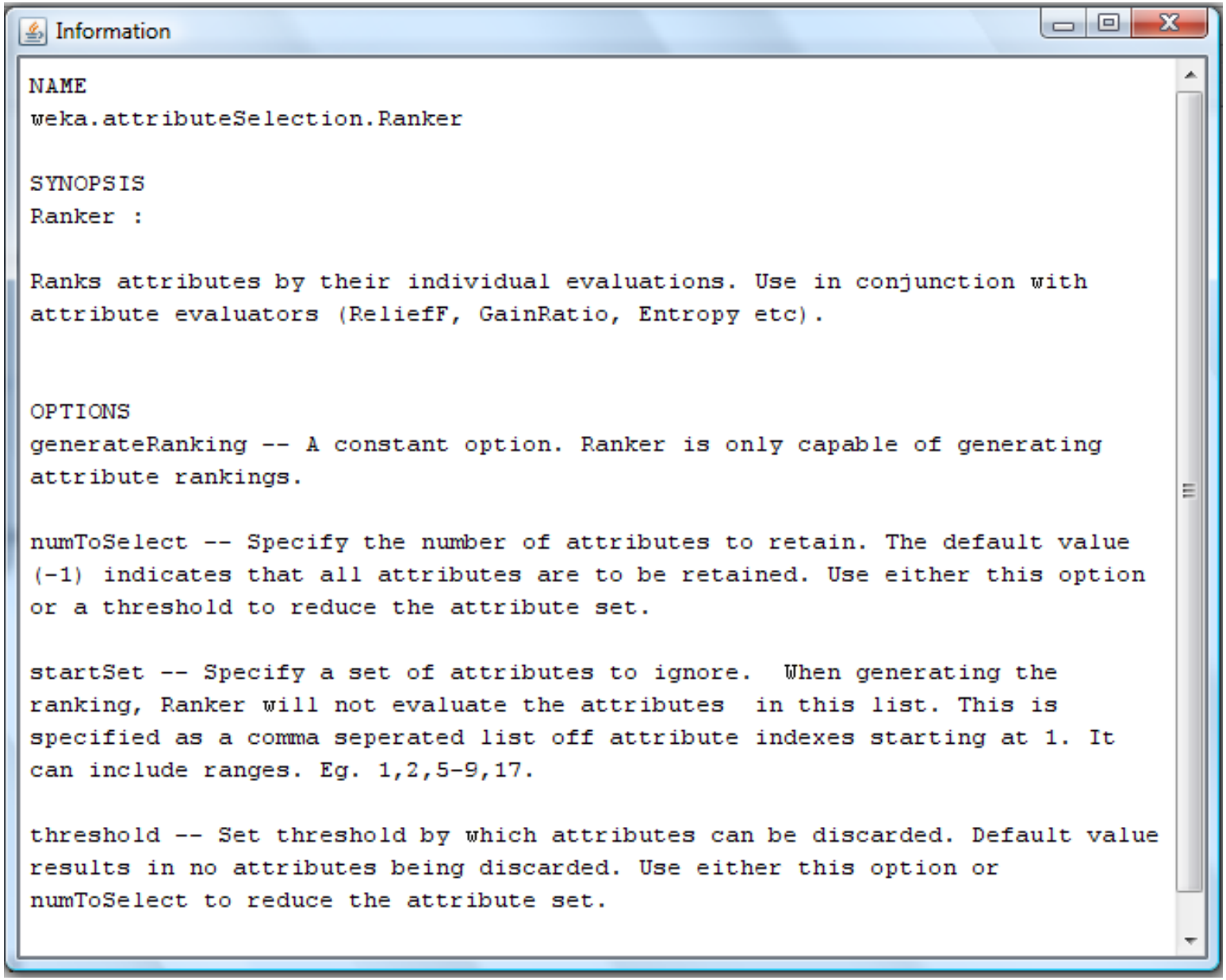
Info


Information
NAME
weka.attributeSelection.GainRatioAttributeEval
SYNOPSIS
GainRatioAttributeEval :
Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.
GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute).
OPTIONS
missingMerge -- Distribute counts for missing values. Counts are distributed across other values in proportion to their frequency. Otherwise, missing is treated as a separate value.

Ranker


weka.gui.GenericObjectEditor
weka.attributeSelection.Ranker
About
Ranker :
More
Ranks attributes by their individual evaluations.
generateRanking
True
numToSelect
-1
startSet
threshold
-1.7976931348623157E308
Open...
Save...
OK
Cancel

Info


Information
NAME
weka.attributeSelection.Ranker
SYNOPSIS
Ranker :
Ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc).
OPTIONS
generateRanking -- A constant option. Ranker is only capable of generating attribute rankings.
numToSelect -- Specify the number of attributes to retain. The default value (-1) indicates that all attributes are to be retained. Use either this option or a threshold to reduce the attribute set.
startSet -- Specify a set of attributes to ignore. When generating the ranking, Ranker will not evaluate the attributes in this list. This is specified as a comma seperated list off attribute indexes starting at 1. It can include ranges. Eg. 1,2,5-9,17.
threshold -- Set threshold by which attributes can be discarded. Default value results in no attributes being discarded. Use either this option or numToSelect to reduce the attribute set.

Results

Attribute selection 10 fold cross-validation (stratified), seed: 1

average merit	average rank	attribute
0.03 +- 0.001	1 +- 0	4 Source
0.018 +- 0	2.1 +- 0.3	3 ZIP3
0.015 +- 0.001	2.9 +- 0.3	9 Triathlete
0.011 +- 0	4 +- 0	7 Miles/Week
0.006 +- 0	5.6 +- 0.49	6 YearsRunning
0.006 +- 0.001	5.7 +- 0.78	8 Races/Year
0.006 +- 0	6.7 +- 0.64	2 State
0.002 +- 0	8 +- 0	5 Age
0.001 +- 0	9 +- 0	1 Gender

Let's Model

- Lessons learned last time:
  - No black-box methods, if not appropriate
  - Explore numeric and nominal classification, if necessary

Build the Models

- Let's create predictive models using the 10-fold evaluation:
  - Representations Tree
  - Decision Tree
  - Grafting Decision Tree
  - other

Representations Tree

- REPTree
- ==========

- ZIP3 = ? : 1 (5/0) [1/0]
- ZIP3 = 100.0
- | Source = RBCK : 2 (0/0) [0/0]
- | Source = ? : 2 (0/0) [0/0]
- | Source = EMC : 2 (9/2) [1/1]
- | Source = EMCD : 1 (1/0) [0/0]
- | Source = DAO3 : 2 (0/0) [0/0]
- | Source = SWP : 1 (3/0) [0/0]
- | Source = DAOD : 1 (2/1) [1/0]
- | Source = DAOA : 2 (1/0) [0/0]
- | Source = DAOL : 2 (0/0) [0/0]
- |...

RepTree: Evaluation

- Size of the tree : 1200

- === Stratified cross-validation ===
- === Summary ===

Correctly Classified Instances	2924	63.9126 %
Incorrectly Classified Instances	1651	36.0874 %
Kappa statistic	0.0524	
Mean absolute error	0.4319	
Root mean squared error	0.501	
Relative absolute error	97.1712 %	
Root relative squared error	106.2729 %	
Total Number of Instances	4575	

Decision Tree

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

- Source = RBCK: 2 (89.0/39.0)
- Source = ?: 2 (24.0/9.0)
- Source = EMC: 1 (2746.0/881.0)
- Source = EMCD: 1 (448.0/191.0)
- Source = DAO3: 1 (318.0/114.0)
- Source = SWP: 1 (405.0/11.0)
- Source = DAOD: 1 (298.0/128.0)
- Source = DAOA: 2 (70.0/34.0)
- Source = DAOL: 2 (32.0/12.0)
- Source = DCF: 2 (17.0)
- Source = ERR: 1 (54.0/14.0)

Decision Tree: Scores

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	3100	67.7596 %
Incorrectly Classified Instances	1475	32.2404 %
Kappa statistic	0.0844	
Mean absolute error	0.4136	
Root mean squared error	0.4568	
Relative absolute error	93.0633 %	
Root relative squared error	96.9119 %	
Total Number of Instances	4575	

Grafting Decision Tree

J48graft pruned tree

Source = RBCK
| State = AZ: 1 (0.0|27.0/1.0)
| State != AZ
| | State = WA: 1 (0.0|23.0/1.0)
| | State != WA
| | | ZIP3 = 930.0: 1 (0.0|8.0)
| | | ZIP3 != 930.0
| | | | ZIP3 = 981.0: 1 (0.0|7.0)
| | | | ZIP3 != 981.0
| | | | | ZIP3 = 949.0: 1 (0.0|7.0)
| | | | | ZIP3 != 949.0
| | | | | | ZIP3 = 875.0: 1 (0.0|7.0)
| | | | | | ZIP3 != 875.0

Grafting DT: Scores

- === Stratified cross-validation ===
- === Summary ===

Correctly Classified Instances	3097	67.694 %
Incorrectly Classified Instances	1478	32.306 %
Kappa statistic	0.0782	
Mean absolute error	0.4141	
Root mean squared error	0.4578	
Relative absolute error	93.1693 %	
Root relative squared error	97.1065 %	
Total Number of Instances	4575	

What's Next:

- Weak scores
- Remove ZIP3?
- Remove State?
- Make an iteration or two and test this path out

RepTree without ZIP3: 66.776 %

REPTree
==========

Source = RBCK
| State = ? : 2 (0/0) [0/0]
| State = MI : 2 (2/0) [1/1]
| State = MA : 2 (0/0) [0/0]
| State = RI : 2 (0/0) [0/0]
| State = NH : 2 (0/0) [0/0]
| State = NY : 1 (2/0) [5/1]
| State = ME : 2 (0/0) [0/0]
| State = VT : 1 (1/0) [0/0]
| State = CT : 2 (0/0) [0/0]
| State = UT : 2 (0/0) [0/0]
...

RepTree without ZIP3, State: 66.9945 %

REPTree
==========

Source = RBCK
| YearsRunning = ? : 1 (15/6) [5/2]
| YearsRunning = 1.0 : 2 (25/9) [15/9]
| YearsRunning = 7.0 : 2 (4/0) [0/0]
| YearsRunning = 5.0 : 2 (5/2) [4/1]
| YearsRunning = 3.0 : 1 (4/1) [0/0]
| YearsRunning = 0.0 : 2 (1/0) [0/0]
| YearsRunning = 10.0 : 2 (8/3) [3/0]
Source = ? : 2 (20/7) [4/2]
Source = EMC
| Miles/Week = ?
| | Races/Year = ?

Dead-end path…

- The results are not very good
- TP, FP rates are not adequate
- Need to backtrack and change something else
- Let's keep the class distribution ratio to 1:1 and get a little bit more data

Subsample with Replacement

- weka.filters.supervised.instance.Resample -B 1.0 -S 1 -Z 50.0
- Will give about 50% of the original dataset, with similar class distributions
- Save it as Shoe5b.arff

DT: Model

Source = RBCK: 2 (270.0/60.0)
Source = ?: 2 (64.0/6.0)
Source = EMC
| ZIP3 = ?: 1 (0.0)
| ZIP3 = 100.0
| | Miles/Week = ?: 2 (0.0)
| | Miles/Week = 6.0
| | | YearsRunning = ?: 1 (2.0)
| | | YearsRunning = 1.0: 2 (16.0/2.0)
| | | YearsRunning = 7.0: 2 (0.0)
| | | YearsRunning = 5.0: 2 (0.0)
| | | YearsRunning = 3.0: 2 (2.0)
| | | YearsRunning = 0.0: 2 (0.0)
| | | YearsRunning = 10.0: 2 (0.0)
| | Miles/Week = 11.0: 1 (7.0)…

DT: Scores

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	11029	83.0122 %
Incorrectly Classified Instances	2257	16.9878 %
Kappa statistic	0.6599	
Mean absolute error	0.2034	
Root mean squared error	0.3577	
Relative absolute error	40.6821 %	
Root relative squared error	71.5362 %	
Total Number of Instances	13286	

DT: TP/FP rates

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.744	0.084	0.897	0.744	0.813	0.897	1
0.916	0.256	0.783	0.916	0.844	0.897	2

=== Confusion Matrix ===

a	b	<-- classified as
4909	1693	a = 1
564	6120	b = 2

Other Models

- Not as good
- w/out ZIP3 – not as good
- w/out ZIP3 and state – not as good

More DT Variations

- w/out ZIP3 – 79.9488%
- w/out ZIP3 and State - 73.521 %

- ZIP and state are redundant!

Analysis

- 83% accuracy accomplished
- More input into the dataset details needed
- ‘Source’ might need work – typos?
- More experimentation with numeric vs. nominal variables could give beneficial outcome
- Try to get it from the company, if not, more data needed

To Do

- Make a few more iterations
- Backtrack, follow your intuition
- Try to get higher evaluation scores with a balanced TP/FP scores

- Report any score higher than 85% correctly classified instances for extra credit (10% of the final grade)

Conclusion

- Real life datasets are unclean and often difficult to model
- One might spend anywhere from 2 or 3 months to more than a year trying to model the data
- Sometimes you just need to wait for more data, or ask for a different protocol for collecting the data, more input variables, etc.
- You will learn tricks of the trade as you go along!

Next

- Next is Lesson 6: Another example of complex datasets and tasks in data mining with the emphasis on meta learning methods