Weka: Data Preprocessing Lab

1. Numeric Cleaner

Open diabetes.arff. Visualize all, attributesplas, pres, mass have missing values

Weka, choose, filter, unsupervised, attribute, numeric cleaner, click, attribute Indices: 6, minDefault: NaN, MinThreshold: 0.1E-7, Ok, Apply, Edit: you can see missing values.

2. Remove missing Values

Go to filter, go to weka, filters, unsupervised, instance, ,removeWithvalues, attributeIndex: 6, matchMissingvalues: True, OK, Apply, Check mass. All missing value records removed.

Undo

3. Impute numeric missing values

Weka, filter, Choose, unsupervised, attribute, replaceMissingValues, Ok, Apply. Ckeck mass. There are no missing values.

4. Discretize

Open credit-g.arff

Select attribute age unsupervised, attribute, Discretize, select on the discretize bar, attribute indices 13 (for age), bins range precision (for decimal values limit) = 2,bins =3, ok, apply, save as type csv

Open file in excel replace values with Old, Middle and Young, save the file as csv

5. Info Gain Attribute Evaluator

Open csv file credit-g-nominal.arff in weka

select attributes from top bar

attribute Evaluator

InfogainAttributeEval

Alert- yes for ranker

Start

Check Results

Select attributes : 17,19,18,8,11,16, remove, save

6. Change any attribute as class

Open mpg.arff

Edit

Select mpg, set attribute as class, ok

7. Change Numeric to Nominal

Open diabetes.arff

Select attribute preg- numeric

Weka, filters, unsupervised, attribute, NumericToNominal, Click on bar, attribute indices 1, Apply

8. Normalize/ Standardize

Open iris.arff, check values of all attributes. Each has a different range.

Weka, filters, unsupervised, attribute, normalize, apply (all values between 0 and 1)

Undo, standardize, apply(mean 0, std dev=1

9. Remove Nominal Attribute Missing values

Open soybean.arff

Select attribute plant-stand. It has missing values

Weka, filters, unsupervised,instance, RemoveWithValues, click bar, attribute indices : 2, invert Selection: true, matchMissingValues: True, OK, Apply

10. Finding and removing Outliers/ Extreme Values (Applicable for file having no missing values only and only numeric attributes)

Open file cpu.arff

Weka, filters, unsupervised, attribute, InterQuartilerange, Apply

Two extra columns added. Edit, Select column outlier, set class as outlier, OK. visualize

Weka,Filters , unsupervised, instance, removeWithvalues, click on bar

Attributeindex: 9

Attribute outlier has two values no(1) and yes(2). We want to remove outliers, so nominal indices=2 or last.. ok, Apply. save as a new file

11. Numeric transform

Iris.arffweka filter unsupervised attribute NumericTransform, attributIndices: 1, metod name : floor

12. PCA

Open file cpu.arff, filter, unsupervised, attribute, PrincipalComponents, click, variance covered:0.95, ok, apply.

Check for variance/Std deviation on the right. It is the maximum variance, Set threshold=50% of the maximum. All other attributes have less than 40%. Select them (4,5) and click remove

When we know how many attributes to keep:

Select Attribute, Attribute Evaluator, principal Components, Search Method: ranker, Click on Components, maximimAttributenames: 5, varianceCovered: 0.95, Ok, Start

You can see all five attributes.

Click on Ranker, numToSelect: 3, OK, Start. It select the best 3 newfeatutres.

13. Training and Test Set

Open iris.arff 150 instances

filter choose weka filter unsupervised, instance, Resample, click for properties of filter,

invertSelection: false,

noReplacement: True,

sampleSizepercent:60,

OK, Apply , 90 instances.   Save as iris_train.arff. Undo

Filter ,choose weka filter, unsupervised, instance, Resample, click for properties of filter,

invertSelection: True,

noReplacement: True,

sampleSizepercent:60,

OK, Apply , 60 instances.   Save as iris_test.arff. Undo

14. Random Undersampling

Open credit-g.argg

Click class , Good :700, bad : 300. Imbalance

Weka ,filter, supervised, instance, spreadSubSample, click,

distributonSpread: 1 (Which value to subsample) ,

Ok, Apply

15. Oversampling

Weka, Tools, package manager, Package search,SMOTE, , Enter, Select SMOTE, install

Weka Explorer

Open credit-g.arff

Click class , Good :700, bad : 300. Imbalance

Weka ,filter, supervised, instance, SMOTE, click, classValue: 2 (Which class value to oversample),

nearesrtNeighbours: 5, Ok, Apply

Check no of instance of class. They have increased by 100% for classValue 2. Edit. All newly inserted records are at the bottom. Randomize them.

Weka, filter, unsupervised, instance, randomize, apply. Check by edit

16. Append/ Merge

Select Weka application SimpleCLI

java weka.core.Instances append d:\iris_train.arff d:\iris_test.arff > d:\iris_total.arff

Enter

Java weka.core.Instances merge d:\iris_train.arff d:\iris_test.arff > d:\iris_merge.arff

Enter

17. Nominal to Binary/Numeric to Binary

Open credit.arff

Filter: Supervised, attribute,Nominal to binary, Apply.

Associate: Start button not enabled

Preprocess

Filter, unsupervised, attribute, numericToBinary, ignoreClass: True

18. Association Rule mining

Apriori requires file with nominal/binary attribute

Open weather nominal.arff

Associate, Apriori

Delta:0.05

LowerBoundMinSupport: 0.2

minMetric: Confidence

OutputIemSets: True

Note: Apriori when used on large databases gives memory error, so use a smaller dataset

FPGrowth: Requires file with binary attributes only

Open weathernominal.arff

Filter: NominalToBinary, Apply

Filter, unsupervised, numericToBinary, Apply

Associate FPGrowth

Classify

open file: credit-g.arff

It has 21 attributes. Classify: j48, start, 70.5% accuracy. Select attribute at top. Attribute Evaluator, Preprocess Tab: Select and Remove 18,8,11,16. Classify: j48, start, check accuracy. Increased to 72%

Choose, InformationainAttributeEval. Search Method ranker, Click, numToSelect: 10, Ok, Start