Contents

SUPERVISED ATTRIBUTE

NAME - weka.filters.supervised.attribute.AttributeSelection

SYNOPSIS
A supervised attribute filter that can be used to select attributes. It is very flexible and allows various search and evaluation methods to be combined.

OPTIONS
evaluator -- Determines how attributes/attribute subsets are evaluated.

search -- Determines the search method.

NAME - weka.filters.supervised.attribute.ClassOrder

SYNOPSIS
Changes the order of the classes so that the class values are no longer of in the order specified in the header. The values will be in the order specified by the user -- it could be either in ascending/descending order by the class frequency or in random order. Note that this filter currently does not change the header, only the class values of the instances, so there is not much point in using it in conjunction with the FilteredClassifier.

OPTIONS
classOrder -- Specify the class order after the filtering

seed -- Specify the seed of randomization of the class order

NAME - weka.filters.supervised.attribute.Discretize

SYNOPSIS
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. Discretization is by Fayyad & Irani's MDL method (the default).

OPTIONS

attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

invertSelection -- Set attribute selection mode. If false, only selected (numeric) attributes in the range will be discretized; if true, only non-selected attributes will be discretized.

makeBinary -- Make resulting attributes binary.

useBetterEncoding -- Uses a more efficient split point encoding.

useKononenko -- Use Kononenko's MDL criterion. If set to false uses the Fayyad & Irani criterion.

NAME - weka.filters.supervised.attribute.NominalToBinary

SYNOPSIS
Converts all nominal attributes into binary numeric attributes. An attribute with k values is transformed into k binary attributes if the class is nominal (using the one-attribute-per-value approach). Binary attributes are left binary, if option '-A' is not given.If the class is numeric, k - 1 new binary attributes are generated in the manner described in "Classification and Regression Trees" by Breiman et al. (i.e. taking the average class value associated with each attribute value into account)

OPTIONS
binaryAttributesNominal -- Whether resulting binary attributes will be nominal.

transformAllValues -- Whether all nominal values are turned into new attributes, not only if there are more than 2.

SUPERVISED INSTANCE

NAME - weka.filters.supervised.instance.Resample

SYNOPSIS
Produces a random subsample of a dataset using sampling with replacement.The original dataset must fit entirely in memory. The number of instances in the generated dataset may be specified. The dataset must have a nominal class attribute. If not, use the unsupervised version. The filter can be made to maintain the class distribution in the subsample, or to bias the class distribution toward a uniform distribution. When used in batch mode (i.e. in the FilteredClassifier), subsequent batches are NOTE resampled.

OPTIONS
biasToUniformClass -- Whether to use bias towards a uniform class. A value of 0 leaves the class distribution as-is, a value of 1 ensures the class distribution is uniform in the output data.

randomSeed -- Sets the random number seed for subsampling.

NAME - weka.filters.supervised.instance.SpreadSubsample

SYNOPSIS
Produces a random subsample of a dataset. The original dataset must fit entirely in memory. This filter allows you to specify the maximum "spread" between the rarest and most common class. For example, you may specify that there be at most a 2:1 difference in class frequencies. When used in batch mode, subsequent batches are NOT resampled.

OPTIONS
adjustWeights -- Wether instance weights will be adjusted to maintain total weight per class.

distributionSpread -- The maximum class distribution spread. (0 = no maximum spread, 1 = uniform distribution, 10 = allow at most a 10:1 ratio between the classes).

maxCount -- The maximum count for any class value (0 = unlimited).

randomSeed -- Sets the random number seed for subsampling.

NAME - weka.filters.supervised.instance.StratifiedRemoveFolds

SYNOPSIS
This filter takes a dataset and outputs a specified fold for cross validation. If you do not want the folds to be stratified use the unsupervised version.

OPTIONS
fold -- The fold which is selected.

invertSelection -- Whether to invert the selection.

numFolds -- The number of folds to split the dataset into.

seed -- the random number seed for shuffling the dataset. If seed is negative, shuffling will not be performed.

UNSUPERVISED ATTRIBUTE

NAME - weka.filters.unsupervised.attribute.Add

SYNOPSIS
An instance filter that adds a new attribute to the dataset. The new attribute will contain all missing values.

OPTIONS
attributeIndex -- The position (starting from 1) where the attribute will be inserted (first and last are valid indices).

attributeName -- Set the new attribute's name.

nominalLabels -- The list of value labels (nominal attribute creation only). The list must be comma-separated, eg: "red,green,blue". If this is empty, the created attribute will be numeric.

NAME - weka.filters.unsupervised.attribute.AddCluster

SYNOPSIS

A filter that adds a new nominal attribute representing the cluster assigned to each instance by the specified clustering algorithm.

OPTIONS
clusterer -- The clusterer to assign clusters with.

ignoredAttributeIndices -- The range of attributes to be ignored by the clusterer. eg: first-3,5,9-last

NAME - weka.filters.unsupervised.attribute.AddExpression

SYNOPSIS
An instance filter that creates a new attribute by applying a mathematical expression to existing attributes. The expression can contain attribute references and numeric constants. Supported opperators are : +, -, *, /, ^, log, abs, cos, exp, sqrt, floor, ceil, rint, tan, sin, (,). Attributes are specified by prefixing with 'a', eg. a7 is attribute number 7 (starting from 1). Example expression : a1^2*a5/log(a7*4.0).

OPTIONS
debug -- Set debug mode. If true then the new attribute will be named with the postfix parse of the supplied expression.

expression -- Set the math expression to apply. Eg. a1^2*a5/log(a7*4.0)

name -- Set the name of the new attribute.

NAME - weka.filters.unsupervised.attribute.AddNoise

SYNOPSIS
An instance filter that changes a percentage of a given attributes values. The attribute must be nominal. Missing value can be treated as value itself.

OPTIONS
percent -- Percentage of introduced noise to data.

randomSeed -- Random number seed.

useMissing -- Flag to set if missing values are used.

NAME - weka.filters.unsupervised.attribute.ChangeDateFormat

SYNOPSIS
Changes the format used by a date attribute.

OPTIONS
attributeIndex -- Sets which attribute to process. This attribute must be of type date ("first" and "last" are valid values)

dateFormat -- The date format to change to. This should be a format understood by Java's SimpleDateFormat class.

NAME - weka.filters.unsupervised.attribute.ClusterMembership

SYNOPSIS
A filter that uses a density-based clusterer to generate cluster membership values; filtered instances are composed of these values plus the class attribute (if set in the input data). If a (nominal) class attribute is set, the clusterer is run separately for each class. The class attribute (if set) and any user-specified attributes are ignored during the clustering operation

OPTIONS
ignoredAttributeIndices -- The range of attributes to be ignored by the clusterer. eg: first-3,5,9-last

NAME - weka.filters.unsupervised.attribute.Copy

SYNOPSIS
An instance filter that copies a range of attributes in the dataset. This is used in conjunction with other filters that overwrite attribute values during the course of their operation -- this filter allows the original attributes to be kept as well as the new attributes.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

invertSelection -- Sets copy selected vs unselected action. If set to false, only the specified attributes will be copied; If set to true, non-specified attributes will be copied.

NAME - weka.filters.unsupervised.attribute.Discretize

SYNOPSIS
An instance filter that discretizes a range of numeric attributes in the dataset into nominal attributes. Discretization is by simple binning. Skips the class attribute if set.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

bins -- Number of bins.

desiredWeightOfInstancesPerInterval -- Sets the desired weight of instances per interval for equal-frequency binning.

findNumBins -- Optimize number of equal-width bins using leave-one-out. Doesn't work for equal-frequency binning

invertSelection -- Set attribute selection mode. If false, only selected (numeric) attributes in the range will be discretized; if true, only non-selected attributes will be discretized.

makeBinary -- Make resulting attributes binary.

useEqualFrequency -- If set to true, equal-frequency binning will be used instead of equal-width binning.

NAME - weka.filters.unsupervised.attribute.FirstOrder

SYNOPSIS

This instance filter takes a range of N numeric attributes and replaces them with N-1 numeric attributes, the values of which are the difference between consecutive attribute values from the original instance. eg:

Original attribute values

0.1, 0.2, 0.3, 0.1, 0.3

New attribute values

0.1, 0.1, -0.2, 0.2

The range of attributes used is taken in numeric order. That is, a range spec of 7-11,3-5 will use the attribute ordering 3,4,5,7,8,9,10,11 for the differences, NOT 7,8,9,10,11,3,4,5.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

NAME - weka.filters.unsupervised.attribute.MakeIndicator

SYNOPSIS
A filter that creates a new dataset with a boolean attribute replacing a nominal attribute. In the new dataset, a value of 1 is assigned to an instance that exhibits a particular range of attribute values, a 0 to an instance that doesn't. The boolean attribute is coded as numeric by default.

OPTIONS
attributeIndex -- Sets which attribute should be replaced by the indicator. This attribute must be nominal.

numeric -- Determines whether the output indicator attribute is numeric. If this is set to false, the output attribute will be nominal.

valueIndices -- Specify range of nominal values to act on. This is a comma separated list of attribute indices (numbered from 1), with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

NAME - weka.filters.unsupervised.attribute.MergeTwoValues

SYNOPSIS
Merges two values of a nominal attribute into one value.

OPTIONS
attributeIndex -- Sets which attribute to process. This attribute must be nominal ("first" and "last" are valid values)

NAME - weka.filters.unsupervised.attribute.NominalToBinary

SYNOPSIS
Converts all nominal attributes into binary numeric attributes. An attribute with k values is transformed into k binary attributes if the class is nominal (using the one-attribute-per-value approach). Binary attributes are left binary.If the class is numeric, you might want to use the supervised version of this filter.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

binaryAttributesNominal -- Whether resulting binary attributes will be nominal.

invertSelection -- Set attribute selection mode. If false, only selected (numeric) attributes in the range will be discretized; if true, only non-selected attributes will be discretized.

NAME - weka.filters.unsupervised.attribute.Normalize

SYNOPSIS
Normalizes all numeric values in the given dataset (apart from the class attribute, if set). The resulting values are in [0,1] for the data used to compute the normalization intervals.

NAME - weka.filters.unsupervised.attribute.NumericToBinary

SYNOPSIS
Converts all numeric attributes into binary attributes (apart from the class attribute, if set): if the value of the numeric attribute is exactly zero, the value of the new attribute will be zero. If the value of the numeric attribute is missing, the value of the new attribute will be missing. Otherwise, the value of the new attribute will be one. The new attributes will nominal.

NAME - weka.filters.unsupervised.attribute.NumericTransform

SYNOPSIS
Transforms numeric attributes using a given transformation method.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

className -- Name of the class containing the method used for the transformation.

invertSelection -- Whether to process the inverse of the given attribute ranges.

methodName -- Name of the method used for the transformation.

NAME - weka.filters.unsupervised.attribute.Obfuscate

SYNOPSIS
An instance filter that obfuscates all strings in the data

NAME - weka.filters.unsupervised.attribute.PKIDiscretize

SYNOPSIS
Discretizes numeric attributes using equal frequency binning, where the number of bins is equal to the square root of the number of non-missing values.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

bins -- Ignored.

desiredWeightOfInstancesPerInterval -- Sets the desired weight of instances per interval for equal-frequency binning.

findNumBins -- Ignored.

invertSelection -- Set attribute selection mode. If false, only selected (numeric) attributes in the range will be discretized; if true, only non-selected attributes will be discretized.

makeBinary -- Make resulting attributes binary.

useEqualFrequency -- Always true.

NAME - weka.filters.unsupervised.attribute.RandomProjection

SYNOPSIS
Reduces the dimensionality of the data by projecting it onto a lower dimensional subspace using a random matrix with columns of unit length (i.e. It will reduce the number of attributes in the data while preserving much of its variation like PCA, but at a much less computational cost).
It first applies the NominalToBinary filter to convert all attributes to numeric before reducing the dimension. It preserves the class attribute.

OPTIONS
distribution -- The distribution to use for calculating the random matrix.
Sparse1 is:
sqrt(3) * { -1 with prob(1/6),
0 with prob(2/3),
+1 with prob(1/6) }
Sparse2 is:
{ -1 with prob(1/2),
+1 with prob(1/2) }

numberOfAttributes -- The number of dimensions (attributes) the data should be reduced to.

percent -- The percentage of dimensions (attributes) the data should be reduced to (inclusive of the class attribute). This NumberOfAttributes option is ignored if this option is present or is greater than zero.

randomSeed -- The random seed used by the random number generator used for generating the random matrix

replaceMissingValues -- If set the filter uses weka.filters.unsupervised.attribute.ReplaceMissingValues to replace the missing values

NAME - weka.filters.unsupervised.attribute.Remove

SYNOPSIS
An instance filter that removes a range of attributes from the dataset.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

invertSelection -- Determines whether action is to select or delete. If set to true, only the specified attributes will be kept; If set to false, specified attributes will be deleted.

NAME - weka.filters.unsupervised.attribute.RemoveType

SYNOPSIS
Removes attributes of a given type.

OPTIONS
attributeType -- The type of attribute to remove.

invertSelection -- Determines whether action is to select or delete. If set to true, only the specified attributes will be kept; If set to false, specified attributes will be deleted.

NAME - weka.filters.unsupervised.attribute.RemoveUseless

SYNOPSIS
Removes constant attributes, along with nominal attributes that vary too much.

OPTIONS
maximumVariancePercentageAllowed -- Set the threshold for the highest variance allowed before a nominal attribute will be deleted.Specifically, if (number_of_distinct_values / total_number_of_values * 100) is greater than this value then the attribute will be removed.

NAME - weka.filters.unsupervised.attribute.ReplaceMissingValues

SYNOPSIS
Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

NAME - weka.filters.unsupervised.attribute.Standardize

SYNOPSIS
Standardizes all numeric attributes in the given dataset to have zero mean and unit variance (apart from the class attribute, if set).

NAME - weka.filters.unsupervised.attribute.StringToNominal

SYNOPSIS
Converts a string attribute (i.e. unspecified number of values) to nominal (i.e. set number of values). You should ensure that all string values that will appear are represented in the first batch of the data.

OPTIONS
attributeIndex -- Sets which attribute to process. This attribute must be a string attribute ("first" and "last" are valid values)

NAME - weka.filters.unsupervised.attribute.StringToWordVector

SYNOPSIS
Converts String attributes into a set of attributes representing word occurrence information from the text contained in the strings. The set of words (attributes) is determined by the first batch filtered (typically training data).

OPTIONS
IDFTransform -- Sets whether if the word frequencies in a document should be transformed into:
fij*log(num of Docs/num of Docs with word i)
where fij is the frequency of word i in document (instance) j.

TFTransform -- Sets whether if the word frequencies should be transformed into:
log(1+fij)
where fij is the frequency of word i in document (instance) j.

attributeNamePrefix -- Prefix for the created attribute names. (default: "")

delimiters -- Set of delimiter characters to use in tokenizing (default: " \n\t.,:'\"()?!"). This option is ignored if onlyAlphabeticTokens option is set to true.

lowerCaseTokens -- If set then all the word tokens are converted to lower case before being added to the dictionary.

normalizeDocLength -- Sets whether if the word frequencies for a document (instance) should be normalized or not.

onlyAlphabeticTokens -- Sets whether if the word tokens are to be formed only from contiguous alphabetic sequences (The delimiter string is ignored if this option is set to true).

outputWordCounts -- Output word counts rather than boolean 0 or 1(indicating presence or absence of a word).

useStoplist -- Ignores all the words that are on the stoplist, if set to true.

wordsToKeep -- The number of words (per class if there is a class attribute assigned) to attempt to keep.

NAME - weka.filters.unsupervised.attribute.SwapValues

SYNOPSIS
Swaps two values of a nominal attribute.

OPTIONS
attributeIndex -- Sets which attribute to process. This attribute must be nominal ("first" and "last" are valid values)

firstValueIndex -- The index of the first value.("first" and "last" are valid values)

secondValueIndex -- The index of the second value.("first" and "last" are valid values)

NAME - weka.filters.unsupervised.attribute.TimeSeriesDelta

SYNOPSIS
An instance filter that assumes instances form time-series data and replaces attribute values in the current instance with the difference between the current value and the equivalent attribute attribute value of some previous (or future) instance. For instances where the time-shifted value is unknown either the instance may be dropped, or missing values used. Skips the class attribute if it is set.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

fillWithMissing -- For instances at the beginning or end of the dataset where the translated values are not known, use missing values (default is to remove those instances)

instanceRange -- The number of instances forward/backward to merge values between. A negative number indicates taking values from a past instance.

invertSelection -- Invert matching sense. ie calculate for all non-specified columns.

NAME - weka.filters.unsupervised.attribute.TimeSeriesTranslate

SYNOPSIS
An instance filter that assumes instances form time-series data and replaces attribute values in the current instance with the equivalent attribute values of some previous (or future) instance. For instances where the desired value is unknown either the instance may be dropped, or missing values used. Skips the class attribute if it is set.

OPTIONS
attributeIndices -- Specify range of attributes to act on. This is a comma separated list of attribute indices, with "first" and "last" valid values. Specify an inclusive range with "-". E.g: "first-3,5,6-10,last".

fillWithMissing -- For instances at the beginning or end of the dataset where the translated values are not known, use missing values (default is to remove those instances)

instanceRange -- The number of instances forward/backward to merge values between. A negative number indicates taking values from a past instance.

invertSelection -- Invert matching sense. ie calculate for all non-specified columns.

UNSUPERVISED INSTANCE

NAME - weka.filters.unsupervised.instance.NonSparseToSparse

SYNOPSIS
An instance filter that converts all incoming instances into sparse format.

NAME - weka.filters.unsupervised.instance.Normalize

SYNOPSIS
An instance filter that normalize instances considering only numeric attributes and ignoring class index.

OPTIONS

LNorm -- The LNorm to use.

norm -- The norm of the instances after normalization.

NAME - weka.filters.unsupervised.instance.Randomize

SYNOPSIS
Randomly shuffles the order of instances passed through it. The random number generator is reset with the seed value whenever a new set of instances is passed in.

OPTIONS
randomSeed -- Seed for the random number generator.

NAME - weka.filters.unsupervised.instance.RemoveFolds

SYNOPSIS
This filter takes a dataset and outputs a specified fold for cross validation. If you want the folds to be stratified use the supervised version.

OPTIONS
fold -- The fold which is selected.

invertSelection -- Whether to invert the selection.

numFolds -- The number of folds to split the dataset into.

seed -- the random number seed for shuffling the dataset. If seed is negative, shuffling will not be performed.

NAME - weka.filters.unsupervised.instance.RemoveMisclassified

SYNOPSIS
A filter that removes instances which are incorrectly classified. Useful for removing outliers.

OPTIONS
classIndex -- Index of the class upon which to base the misclassifications. If < 0 will use any current set class or default to the last attribute.

classifier -- The classifier upon which to base the misclassifications.

invert -- Whether or not to invert the selection. If true, correctly classified instances will be discarded.

maxIterations -- The maximum number of iterations to perform. < 1 means filter will go until fully cleansed.

numFolds -- The number of cross-validation folds to use. If < 2 then no cross-validation will be performed.

threshold -- Threshold for the max allowable error when predicting a numeric class. Should be >= 0.

NAME - weka.filters.unsupervised.instance.RemovePercentage

SYNOPSIS
A filter that removes a given percentage of a dataset.

OPTIONS
invertSelection -- Whether to invert the selection.

percentage -- The percentage of the data to select.

NAME - weka.filters.unsupervised.instance.RemoveRange

SYNOPSIS
A filter that removes a given range of instances of a dataset.

OPTIONS
instancesIndices -- The range of instances to select. First and last are valid indexes.

invertSelection -- Whether to invert the selection.

NAME - weka.filters.unsupervised.instance.RemoveWithValues

SYNOPSIS
Filters instances according to the value of an attribute.

OPTIONS
attributeIndex -- Choose attribute to be used for selection (default last).

invertSelection -- Invert matching sense.

matchMissingValues -- Missing values count as a match. This setting is independent of the invertSelection option.

modifyHeader -- When selecting on nominal attributes, removes header references to excluded values.

nominalIndices -- Range of label indices to be used for selection on nominal attribute. First and last are valid indexes.

splitPoint -- Numeric value to be used for selection on numeric attribute. Instances with values smaller than given value will be selected.

NAME - weka.filters.unsupervised.instance.Resample

SYNOPSIS
Produces a random subsample of a dataset using sampling withreplacement. The original dataset must fit entirely in memory. The number of instances in the generated dataset may be specified.

OPTIONS
randomSeed -- The seed used for random sampling.

sampleSizePercent -- Size of the subsample as a percentage of the original dataset.

NAME - weka.filters.unsupervised.instance.SparseToNonSparse

SYNOPSIS
An instance filter that converts all incoming sparse instances into non-sparse format.