

Rapid Miner

1) Using Rapidminer tool apply the following pre processing technique

a. Normalization techniques

Description :

The normalize operator normalizes the value of the selected attributes. It is used to scale values so they fit in a specific range.

Dataset (Input):

The input ExampleSet is a collection of ages of individual belonging to categories Kid, youth and adult.

Output

The ExampleSet that with selected attributes is normalised. Here the age value is normalised.

b) Aggregation

Description :

The Aggregate operator creates a new ExampleSet from the input ExampleSet showing the results of the selected aggregation functions. Many aggregation functions are supported including SUM, COUNT, MIN, MAX, AVERAGE, GROUP BY.

Input :

Dataset : Age dataset.

Output of other operators can also be used as input.

Group-by-attributes

This operator can group examples of the input ExampleSet into smaller groups using this parameter.

Output.

Group-by-attribute is applied on the category attribute. The output is adult, kid and youth.

c. Data Cleansing :

Different data cleansing techniques are :

1) Replacing Missing Values

The operator replaces missing values in Examples of selected Attributes by a specified replacement. Missing values can be replaced by the minimum, maximum or average value of that Attribute. Zero any value can also be specified.

2) Replace Infinite values

This operator replaces positive or negative values by specified replacements such as none, zero, max-byte, max-int, max-double and missing.

3. Remove Duplicates

The operator removes duplicate examples from an ExampleSet by comparing all examples with each other on the basis of the specified attributes.

4. Input

Dataset : Age Dataset with missing values in age attribute.

Output :

The output is an exampleSet with the missing values replaced (filled) with the average value of the age column.

d) Sampling

Description.

This operator creates a sample from an ExampleSet by selecting examples randomly. The size of a sample can be specified on absolute, relative and probability basis.

Input.

Dataset : Age dataset.

Output :

A randomized sample of the input ExampleSet.

Parameter :

Absolute : The required number of examples is specified in the sample size parameter.

2. Association Mining

Description:

This operator generates a set of association rules from the given set of frequent itemsets.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships.

Design:

Input:

The example set used here is iris dataset.

Discretize by Frequency

This operator converts the selected numerical attributes into nominal attributes by discretizing the numerical attribute into a user-specified number of bins. Bins of equal frequency are automatically generated, the range of different bins may vary.

Nominal to Binominal

This operator changes the type of selected attributes to binominal and also maps all values of these attributes to binominal values i.e. true or false.

FP-Growth

This operator efficiently calculates all frequently-occurring itemsets in an ExampleSet, using the FP-tree data structure.

Create Association Rules

This operator generates a set of association rules from the given set of frequent itemsets.

Output.

- Itemsets : The itemsets that was given a inputs is passed without changing to the output through this port.
- Rules : The association rules are generated