

[Documentation](#) / [Studio](#) / [Operators](#) / Generalized Sequential Patterns



Generalized Sequential Patterns

(RapidMiner Studio Core)

Synopsis

This operator searches sequential patterns in a set of transactions using the GSP (Generalized Sequential Pattern) algorithm. GSP is a popular algorithm used for sequence mining.

Description

This operator searches sequential patterns in a set of transactions. The ExampleSet must contain one attribute for the time and one attribute for the customer. Moreover, each transaction must be encoded as a single example. The time and customer attributes are specified through the *time attribute* and *customer id* parameters respectively. This pair of attributes is used for generating one sequence per customer containing every transaction ordered by the time of each transaction. The algorithm then searches sequential patterns in the form of: If a customer bought item 'a' and item 'c' in one transaction, he bought item 'b' in the next. This pattern is represented in this form: <a, c> then . The minimal support describes how many customer must support such a pattern for regarding it as frequent. Infrequent patterns will be dropped. A customer supports such a pattern, if there are some parts of his sequence that includes that pattern. The above pattern would be supported by a customer, for example, with transactions: <s, g> then <a, s, c> then then <f, h>. The minimum support criteria is specified through the *min support* parameter.

The *min gap*, *max gap* and *window size* parameters determine how transactions are handled. For example, if the above customer forgot to buy item 'c', and had to return 5 minutes later to buy it, then his transactions would look like: <s, g> then <a, s> then <c> then then <f, h>. This would not support the pattern <a, c> then . To avoid this problem, the window size determines, how long a subsequent transaction is treated as the same transaction. If the window size is larger than 5 minutes then <c> would be treated as being part of the second transaction and hence this customer would support the above pattern. The *max gap* parameter causes a customers sequence not to support a pattern, if the transactions containing this pattern are too widely separated in time. The *min gap* parameter does the same if they are too near.


This technique overcomes some crucial drawbacks of existing mining methods, for example:

- absence of time constraints: This drawback is overcome by the min gap and max gap parameters.
- rigid definition of a transaction: This drawback is overcome by the sliding time window.

Please note that all attributes (except customer and time attributes) of the given ExampleSet should be binominal, i.e. nominal attributes with only two possible values. If your ExampleSet does not satisfy this condition, you may use appropriate preprocessing operators to transform it into the required form. The discretization operators can be used for changing the value of numerical attributes to nominal attributes. Then the Nominal to Binominal operator can be used for transforming nominal attributes into binominal attributes.


Please note that the sequential patterns are mined for the positive entries in your ExampleSet, i.e. for those nominal values which are defined as positive in your ExampleSet. If your data does not specify the positive entries correctly, you may set them using the *positive value* parameter. This only works if all your attributes contain this value.

Input

-  example set (*Data Table*)

This input port expects an ExampleSet. Please make sure that all attributes (except customer and time attributes) of the ExampleSet are binominal.

Output

-  example set (*Data Table*)

The ExampleSet that was given as input is passed without changing to the output through this port. This is usually used to reuse the same ExampleSet in further operators or to view the ExampleSet in the Results Workspace.

-  patterns

The GSP algorithm is applied on the given ExampleSet and the resultant set of sequential patterns is delivered through this port.

Parameters

- **customer_id**

This parameter specifies the name of the attribute that will be used for identifying the customers.

Range: string

- **time_attribute**

This parameter specifies the name of the numerical attribute that specifies the time of a transaction.

Range: string

- **min_support**

Prune patterns that are supported by less than *min support* percentage of the customers.

Range: real

- **window_size**

The time window within successive transactions will be additionally handled as a single transaction.

Range: real

- **max_gap**

The *max gap* parameter causes a customers sequence not to support a pattern, if the transactions containing this pattern are too widely separated in time.

Range: real

- **min_gap**

The *min gap* parameter causes a customers sequence not to support a pattern, if the transactions containing this pattern are too near in time.

Range: real

- **positive_value**

This parameter determines which value of the binominal attributes should be treated as positive. The attributes with this value in an example are considered to be part of that transaction.

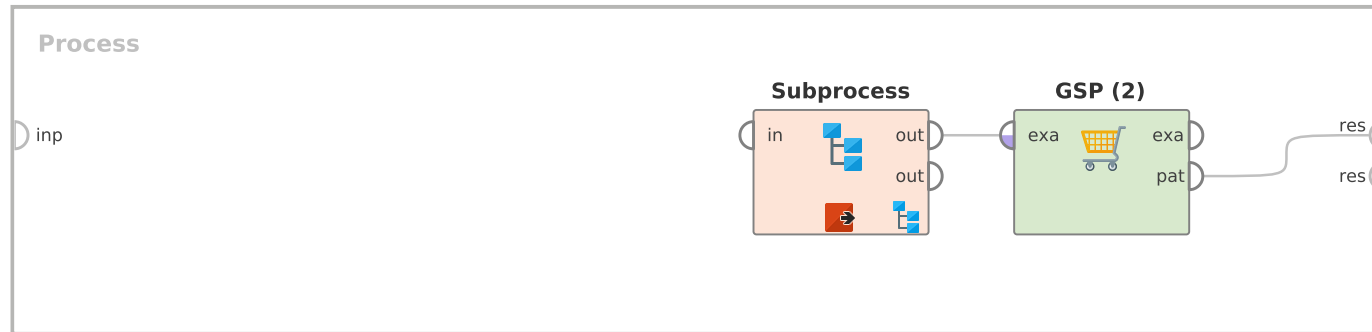
Range: string

Tutorial Processes

Introduction to the GSP operator

The ExampleSet expected by the GSP operator should meet the following criteria: It should have an attribute that can be used for identifying the customers. It should have a numerical attribute that represents the time of the transaction. All other attributes are used for representing items of transactions. These attributes should be binominal.

This Example Process starts with the Subprocess operator. A sequence of operators is applied in the subprocess to generate an ExampleSet that satisfies all the above mentioned conditions. A breakpoint is inserted after the Subprocess operator so that you can have a look at the ExampleSet. The Customer attribute represents the customers, this ExampleSet has five. The Time attribute represents the time of transaction. For simplicity the Time Attribute consists of 20 days. In real scenarios unix time should be used. There are 20 binominal attributes in the ExampleSet that represent items that the customer may buy in a transaction. In this ExampleSet, value 'true' for an item in an example means that this item was bought in this transaction (represented by the current example). The GSP operator is applied on this ExampleSet. The customer id and time attribute parameters are set to 'Customer' and 'Time' respectively. The positive value parameter is set to 'true'. The min support parameter is set 0.9. The resultant set of sequential patterns can be seen in the Results Workspace.



© RapidMiner GmbH 2017. All rights reserved.