Transaction Clustering

Description

Clustering is based on the **CLOPE algorithm**. Its usage enables to process huge arrays of **transactional data**: checks in supermarkets, tracking data of web resources, etc. It is required to provide clustering of the whole set of transactions to place all similar transactions to one cluster, whereas the transactions that differ from each other must be placed into different clusters.

The algorithm selects the number of clusters automatically. An analyst can have influence on the result by means of repulsion coefficient and determination of the maximum number of clusters or iterations.

To get resulting data sets, it is required to provide preliminary training of the handler.

Ports

Input

- — Input data source (data table). Optional.
- — Supplementary input (data table). Optional.

Input data means the transactions generated in 2 fields one of which is a transaction, whereas the second one is an item. For example, the first field is the buyer's code, the second one - a list of his purchases.

Requirements to the Received Data

The fields with the continuous data kind are not suitable for processing.

It is always possible to convert the standard data set that consists of columns with measured object properties to the required kind (with the "Transaction" and "Item" fields) using such Megaladata handling nodes as Unpivoting and Calculator.

Output

- — **Clustering.**

The data table that consists of:

- The field with the "Transaction" usage type.
- Numbers of clusters: each transaction is assigned with the number of the cluster into which it is included.

- — **Cluster parameters.**

The data table that consists of:

- Cluster numbers: numbers of the generated clusters are listed.
- N — count of the transactions included into cluster.
- W — cluster width, namely, the number of the unique objects included into it.
- S — cluster size (area).

Wizard

At this stage it is required to provide the following settings:

- **Fields for clustering.**
  - It is required to set the "Transaction" and "Item" usage types for the fields that are included into processing.
  - "Unspecified" is preserved for other fields.
- **Repulsion coefficient**: it enables to control the level of transactions similarity inside the cluster and consequently the final number of clusters. The higher coefficient, the lower the similarity level and the more clusters will be generated. The default repulsion coefficient value is equal to 2.6. Range of values change varies from 1 to 4.
- **Limit the number of clusters** enables to set the highest number of clusters that can be finally provided by the algorithm manually. This setting can be used if the definite number of cluster is required, for example, not exceeding 15.
- **Limit the number of iterations** enables to terminate the algorithm operation deliberately if the algorithm iterations number exceeds the set maximum number of iterations. The first transaction table check in the algorithm is required for the first clustering defined by the cost function. Then, to improve the clustering quality and optimize the function, the tables are additionally scanned once again several times, until changes in clustering are stopped. It is required to limit the maximum number of iterations in the case of large amount of data for preliminary evaluation of the clustering quality.