The ASSOC Procedure

Overview

Procedure Syntax
   PROC ASSOC Statement
   CUSTOMER Statement
   TARGET Statement

Details

Example

References

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
Overview

Association discovery is the identification of items that occur together in a given event or record. This technique is also known as market basket analysis. Online transaction processing systems often provide the data sources for association discovery. Associations rules are based on frequency counts of the number of times items occur alone and in combination in the database. The rules are expressed as "if item A is part of an event then item B is also part of the event X percent of the time." The rules should not be interpreted as a direct causation but as an association between two or more items. Identifying creditable associations can help the business technologist make decisions such as when to distribute coupons, when to put a product on sale, or how to layout items in a store.

Hypothetical association discovery rules include: If a customer buys shoes, then 10% of the time he also buys socks. A grocery chain may find that 80% of all shoppers are apt to buy a jar of salsa when they also purchase a bag of tortilla chips. When "do-it-yourselfers" buy latex paint they, also buy rollers 85% of the time. Forty percent of investors holding an equity index fund will have a growth fund in their portfolio.

An association rule has a left side (antecedent) and a right side (consequent). Both sides of the rule can contain more than one item. The confidence factor, level of support, and lift are three important evaluation criteria of association discovery. The strength of an association is defined by its confidence factor, which is the percentage of cases in which a consequent appears with a given antecedent. The level of support is how frequently the combination occurs in the market basket (data base). Lift is equal to the confidence factor divided by the expected confidence. A creditable rule has a large relative confidence factor, a relatively large level of support, and a value of lift greater than 1. Rules having a high level of confidence but little support should be interpreted with caution.

The maximum number of items in an association determines the maximum size of the item set to be considered. For example, the default of 4 items indicates that up to 4-way associations are performed.
Procedure Syntax

PROC ASSOC <option(s)>;

CUSTOMER variable-list;

TARGET variable;

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
The ASSOC Procedure

PROC ASSOC Statement

Invokes the ASSOC procedure.

PROC ASSOC <option(s)>;

Required Argument

OUT=libref.SAS-data-set
Specifies the output data set that contains the following variables: SET_SIZE, COUNT, ITEM1, ITEM2,...ITEMn (where n is the maximum number of variables). See Details for more information.

SET_SIZE:
Variable that contains the total number of transactions in the data set. The first observation has the SET_SIZE equal to 0. SET_SIZE is labeled as Relations in the Results Browser.

COUNT:
Contains the number of transactions meeting the rule.

ITEM1, ITEM2,...ITEMn:
Contains the individual items forming the rule including the arrow.

Tip: The OUT= data set created by PROC ASSOC is input to the RULEGEN and SEQ procedures. Run PROC ASSOC and PROC RULEGEN to perform association discovery. Run PROC ASSOC and PROC SEQ to perform sequence discovery.

Options

DATA=libref.SAS-data-set
Identifies the input data source. To perform association discovery, the input data set must have a separate observation for each product purchased by each customer. You must also assign the ID model role to a variable and the TARGET model role to another variable in the Input Data Source.

DMDBCAT=libref.SAS-catalog
Identifies the data catalog of the input data source.

ITEMS=integer
Specifies the maximum number of events or transactions to chain (or associate) together.

SUPPORT=integer
Specifies the minimum number of transactions that must be considered in order for a rule to be
accepted. Rules that do not meet this support level are rejected. The level of support represents how frequently the combination occurs in the market basket (input data source).

| Default: | 5% of the largest item frequency count |

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
CUSTOMER Statement

Specifies the customer(s) to be analyzed.

Alias: CUST

CUSTOMER variable-list;

Required Argument

variable-list

Specifies one or more names of customers to be analyzed.
TARGET Statement

Specifies the target to be analyzed.

TARGET variable;

Required Argument

variable

Specifies the NOMINAL variable, which contains items usually ordered by customers.

Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.
Details

The input to the ASSOC procedure has the following role variables: ID and TARGET. All records with the same ID values form a transaction. Every transaction has a unique ID value and one or more TARGET values.

You may have more than one ID variable. However, associations analysis can only be performed on one target variable at a time. When there are multiple ID variables, PROC ASSOC concatenates them into a single identifier value during computation.

For numeric target variables, missing values constitute a separate item or target level and show up in the rules as a period (.). For character target variables, completely blank values constitute a separate item (target level) and show up in the rules as a period (.). All records with missing ID values are considered a single valid transaction.

Output Processing

PROC ASSOC makes a pass through the data and obtains transaction counts for each item. It outputs these counts with a SET_SIZE of 1 and the items listed under ITEM1. Items that do not meet the support level are discarded. By default, the support level is set to 5% of the largest item count.

PROC ASSOC then generates all potential 2-item sets, makes a pass through the data and obtains transaction counts for each of the 2-item sets. The sets that meet the support level are output with SET_SIZE of 2 and items listed under ITEM1 and ITEM2.

The entire process is repeated for up to n-item sets. The output from PROC ASSOC is saved as SAS data sets. The data sets enable you to define your own evaluation criteria and/or reports.

Note that the ordering of n-items within an n-item set is not important. Any individual transaction, where each of the n-items occurs in any order, qualifies for a count to that particular set. The support level, once set, remains constant throughout the process.

Caution: The theoretical potential number of item sets can grow very quickly. For example, with 50 different items, you have 1225 potential 2-item sets and 19,600 3-item sets. With 5,000 items, you have over 12 million of the 2-item sets, and a correspondingly large number of 3-item sets.

Processing an extremely large number of sets could cause your system to run out of disk and/or memory resources. However, by using a higher support level, you can reduce the item sets to a more manageable number.
Example

PROC ASSOC must be executed before PROC RULEGEN or PROC SEQUENCE is run.

Please see the RULEGEN and SEQUENCE procedures syntax for examples of PROC ASSOC code.
References


Copyright 2000 by SAS Institute Inc., Cary, NC, USA. All rights reserved.