

ECT7110 E-Commerce Data Mining Technique SAS Enterprise Miner Tutorial (Clustering)

I. Clustering Node

Some background:

- cluster similar data into groups

In EM, **K-means Clustering**

1. Initialization

- Set no. of cluster (K) in the entire data set
- Set K initial points to be the means of initial clusters (initial seeds)

2. Refine Clusters (iterative)

- For each data in the data set, calculate its distance between the current clusters
- Assign each data to its nearest cluster (some methods use other criterion to assign data to cluster).
- Recalculate the cluster info. (mean, distribution, ...)

Repeat Process 2 until the K clusters become stable.

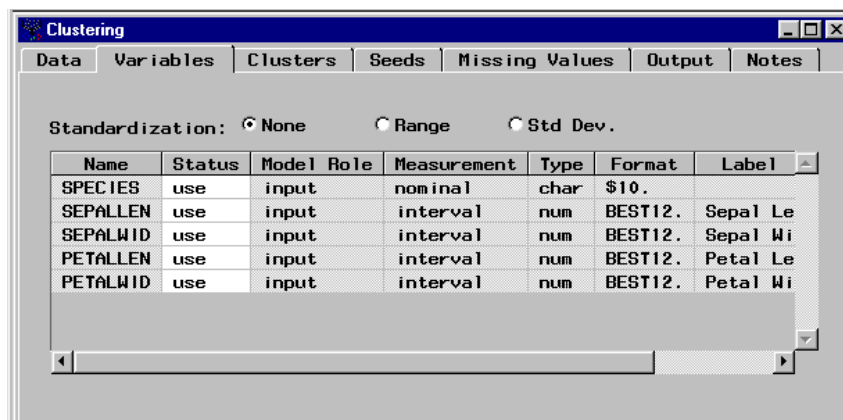
N.B. Parameters need to adjusted:

1. no. of clusters
2. when evaluating the quality of cluster setting, you usually need to calculate distances between clusters → distance measure for clusters
3. distance measure between data and cluster → in EM, Euclidean distance is used:

$$dist(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2}$$

4. clustering criterion to assign data to cluster in the iterative process

Variables Tab

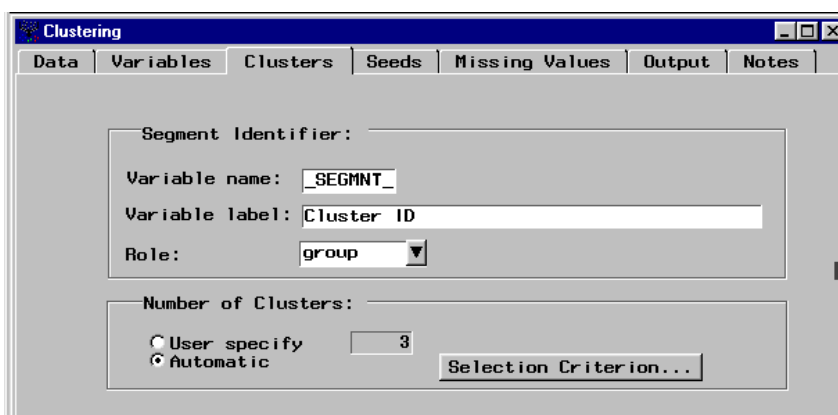


Standardization:

None (default)	not standardized
Range	divided by the range
Std Dev.	divided by the standard deviation

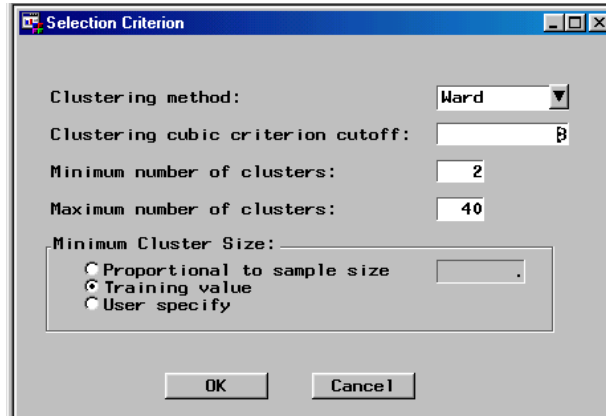
Clusters Tab

- specify options for the *Segment Identifier* and to specify the *Number of Clusters*. The term segment (profile segment) refers to a cluster of observations



Number of Clusters

- User specify
- Automatic: press [*Selection Criterion*] button to change the way the tool automatically determines the number of clusters:



Clustering method: method to define distance between two clusters

Average	Average distance between points in the two clusters
Centroid	Distance between two cluster means
Ward (default)	$\frac{\text{Centroid}_1}{\text{no. of data in cluster 1}} + \frac{\text{Centroid}_2}{\text{no. of data in cluster 2}}$

Clustering cubic criterion cutoff (default is 3):

1. clustering cubic criterion (CCC): method to evaluate the quality of cluster
2. cluster setting with CCC > cutoff will be chosen

Minimum number of clusters (default is 2)

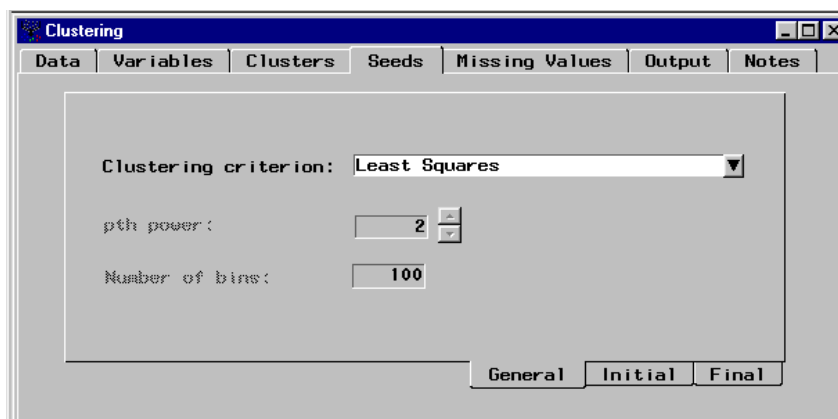
Maximum number of clusters (default is 40)

Minimum Cluster Size

1. *Proportional to sample size*:
 e.g. if
 1. no. of training data = 10000
 2. *Minimum Cluster Size* in *Final* subtab = 500
 3. *Proportional to sample size* = 2000
 then

$$\text{Minimum Cluster Size} = (2000 * [500 / 10000]) = 100$$
2. *Training value*:

$$\text{Minimum Cluster Size} = 500$$
3. *User specify*

Seeds TabGeneral Subtab

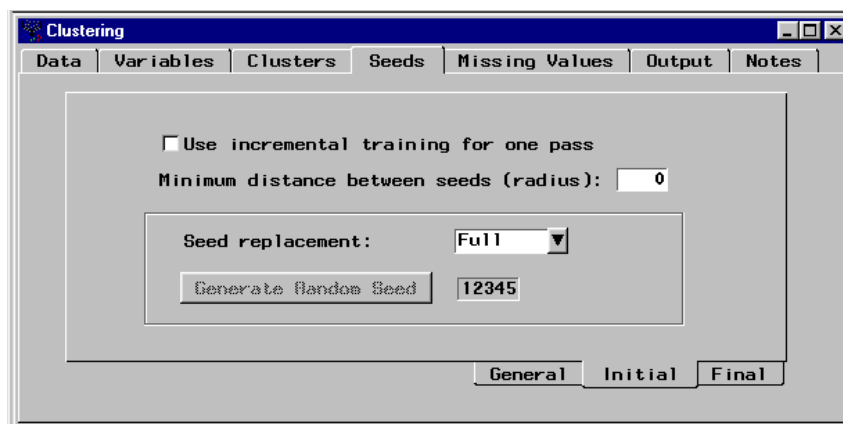
Clustering criterion:

- given a data set, you have criterion used to construct clusters

Mean Absolute Deviation (Median)	Deviation
Modified Eklblom-Newton	Deviation
Least Squares (Fast)	Minimizes the sum of squared distance of data to the cluster mean, only one iteration is performed
Least Squares	Same as above with more than one iteration
Newton	Deviation
Midrange	minimizes the midrange distance of the data points from the cluster means

Initial Subtab

- specify how the cluster seeds are to be initialized

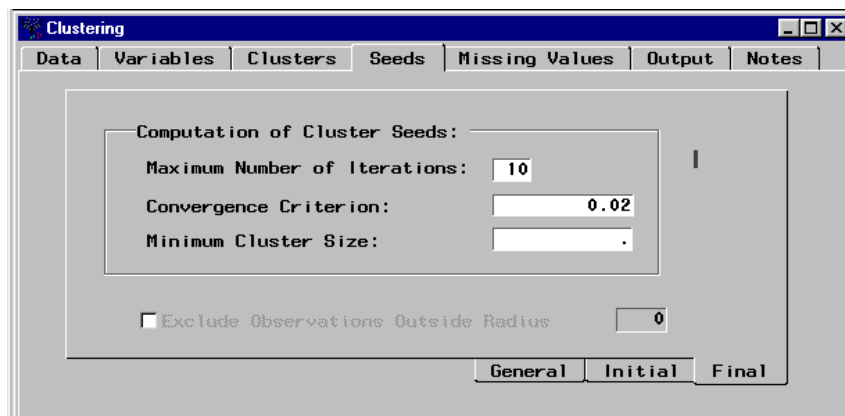
*Use incremental training for one pass:*

- the seeds are allowed to drift as the algorithm selects initial seeds

Seed replacement:

- specify initial seeds (initial seeds must be complete - contains no missing values)

Full	Seeds are chosen to be as far apart as possible
Partial	Only seeds not meet minimum distance requirement will be replaced
None	Choose first n complete data in the data set
Random	Randomly select complete cases

Final Subtab*Convergence Criterion:*

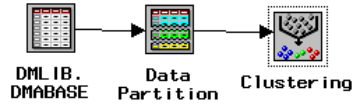
- non-negative

- iterations terminate when the maximum distance by which any seed has changed is less than or equal to the minimum distance between initial seeds times this value

Missing Values Tab

- specify how data observations containing some missing values are to be handled

An Example of Clustering



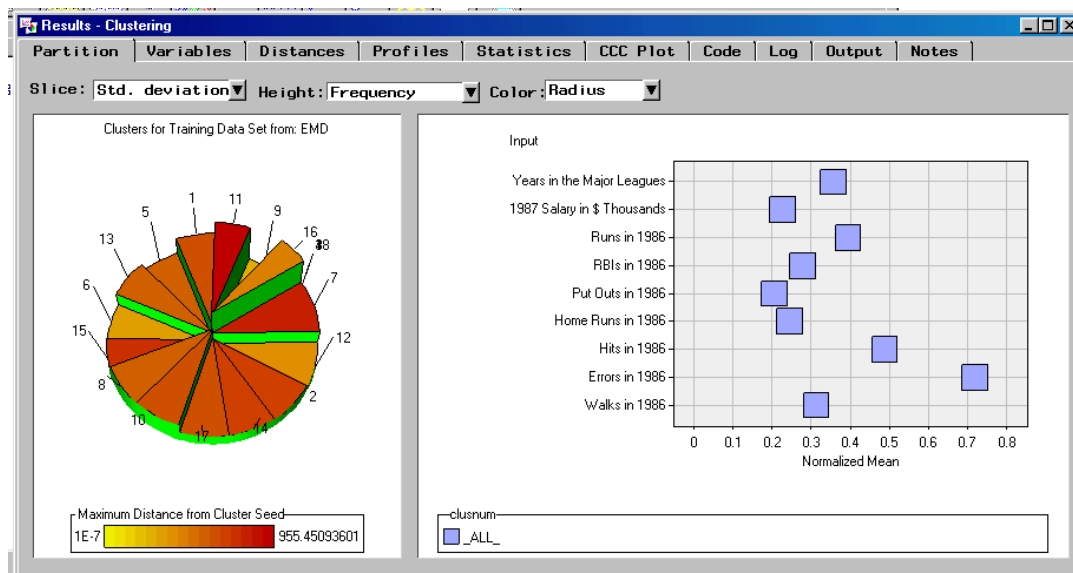
Data set : **SAMPSIO.DMABASE** (no need to set target variable)
 Data Partition : Random; Training: 70%; Validation: 30%
 Clustering : default

Running the Clustering Node → View results

Results Browser

Partition Tab

- provides a graphical representation of key characteristics of the clusters



Left: three-dimensional pie chart with the following default settings:

1. Std deviation (SD of instances in cluster)
2. Frequency (no. of instances in cluster)
3. Radius (size of cluster → distance of instance farthest from the cluster mean)

Right: A grid plot of the input means for the entire training data set

Right-Click Pop-up Menu of Left Pie Chart

Select clusters to view...

- enables you to select clusters, a sorting variable, and a sorting order.
- In the first run of the Clustering node, the pie chart may be difficult to interpret because there are too many segments. The pie chart may be much easier to interpret when a subset of the segments are selected for display

Summary statistics...

- shows summary statistics for interval and class variables of data in the selected cluster

Cluster Profile...

- displays a decision tree (path) for one selected cluster
- enables you to identify influential inputs (variables)

View data...

- displays a data table for the selected clusters

Save as data set...

- enables you to save the observations for the selected cluster(s) to a data set from *Cluster Data Set* window
- you can use a successor node to read and process the cluster subset data set

Delete subset data sets...

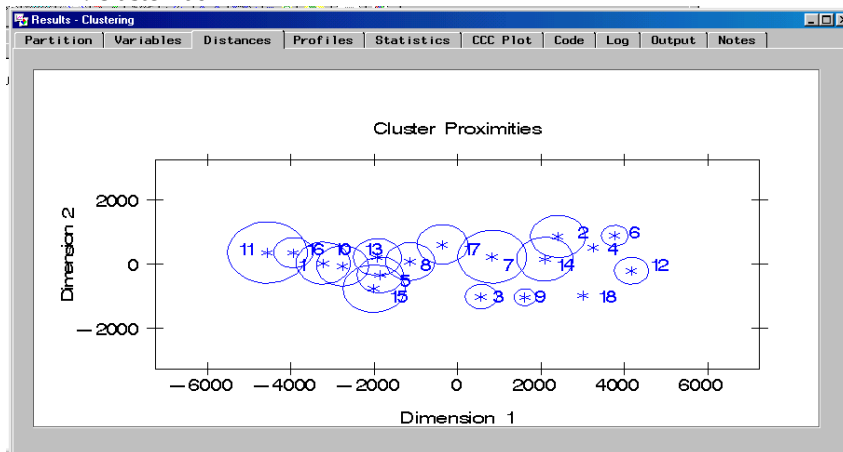
- delete the subset data sets that you created in the *Cluster Data Set* window

Variables Tab

- lists all of the input variables that are used in the Clustering analysis
- *Importance* value is computed to represent the relative importance of the variable to the formation of the clusters (1: most important)

Distances Tab

- provides a graphical representation of the size of each cluster and the relationship among clusters
- axes are determined from multidimensional scaling analysis
- asterisks : cluster centers
- circles : cluster radii



- To display a table of cluster distances select:

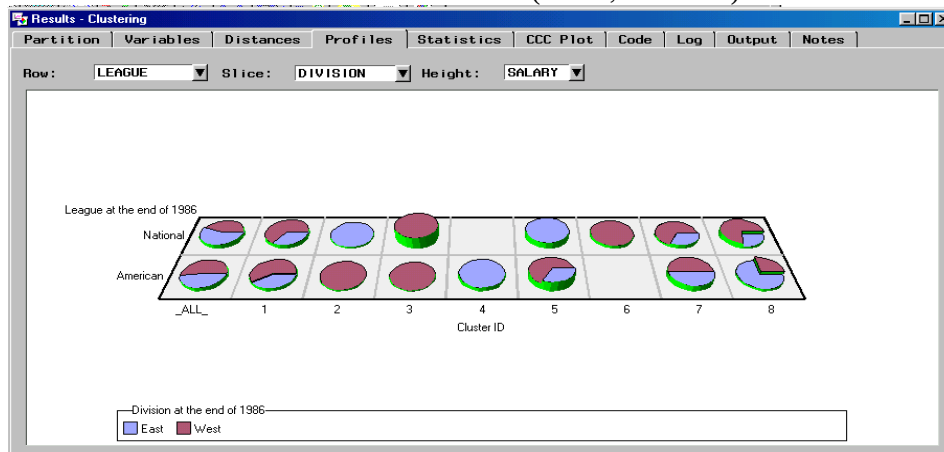
Pull down menu: Format → Distance Table

Profiles Tab

- provides a graphical representation of the Categorical Variables and the Interval Variables for each cluster
- to swap viewing Categorical and Interval variables:
pull down menu *View* → *Categorical / Interval Variables*

Categorical Variables

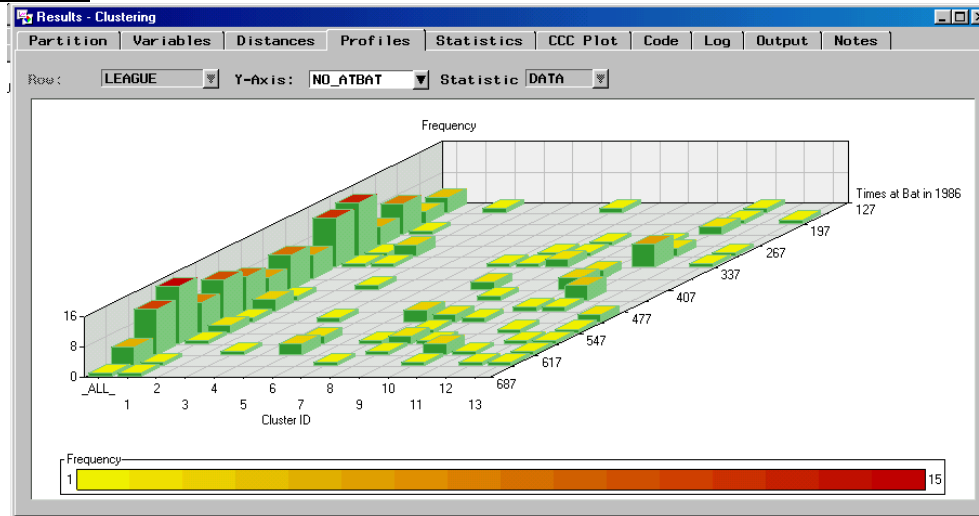
- pull down menu *Format* → *Set Grid dimensions...*(row: 2; column: 9)



- The row variable displays one categorical variable

- The slice variable displays a second categorical variable
- The height variable displays an interval variable

Interval Variables



Statistics Tab

- displays information about each cluster in a tabular format

Frequency of the Cluster	number of cases in each cluster
Root-Mean Square Standard Deviation	the root mean square error across variables of the cluster standard deviations
Maximum Distance from Cluster Seed	the maximum distance from the cluster seed to any case in the cluster
Nearest Cluster	the number of the cluster with a mean closest to the mean of the current cluster
Distance to Nearest Cluster	the distance between the centroids (means) of the current cluster and the nearest other cluster
(Remaining columns)	Mean of each Input Variable

- View *Statistics Plot*
Pull down menu View → *Statistics Plot*

CCC Plot Tab

- displays a plot of the *Cubic Clustering Criterion* plotted against the number of clusters that the Clustering node automatically selects
- activate only if *Automatic in the Number of Clusters* check box in the *Clusters* tab is selected

ECT7110 E-Commerce Data Mining Technique SAS Enterprise Miner Tutorial (Association rule)

Some background:

1. Association Discovery

- the identification of items that occur together in a given event or record
- also known as market basket analysis
- based on frequency counts of the number of times items occur alone and in combination in the database
- expressed as “If (antecedent), then (consequent)”
- e.g. If a customer buys shoes, then 10% of the time he also buys socks.
- Data Layout for Association Discovery

Customer (id)	Product (target)
---------------	------------------

Skip Hops	Natural Choice Tomato Soup
Skip Hops	Washington Carver Honey Ale
Mary Lamb	Natural Choice Black Bean Soup
Mary Lamb	Otto's Oatmeal

2. Sequence Discovery

- goes one step further than association discovery by taking into account the ordering of the relationships among items (the rules additionally imply a timing element)
- rule $A \rightarrow B$ implies that event B occurs after event A occurs
- Of those customers who currently hold an equity index fund in their portfolio, 15% of them will open an international fund in the next year
- apply a sequence model role to a time stamp variable for timing comparisons
- Data Layout for Sequence

Customer (id)	Visit (sequence)	Product (target)
---------------	------------------	------------------

Skip Hops	1	Natural Choice Tomato Soup
Skip Hops	1	Brown Cow 2% Milk
Skip Hops	2	Lucky Dog Chow
Mary Lamb	1	Natural Choice Black Bean Soup
Mary Lamb	2	Washington Carver Honey Ale
Mary Lamb	2	Otto's Oatmeal

3. Evaluation Criteria

Let:

Total transactions	1, 000, 000
Shoes	200, 000
Socks	50, 000
Shoes and Socks	20, 000

For the rule: if buy shoes, then 10% will buy socks

Confidence factor (C)	20,000/200,000 = 10%
Level of support	20,000/1,000,000 = 2%
Expected confidence (EC)	50,000/1,000,000 = 5%
Lift	C/EC = 10%/5% = 2

- A creditable rule has a large confidence factor, a 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.

General Tab

Association

Time Constraints | Sort | Output | Selected Output | Notes

Data | Variables | General | Processing Options | Sequences

Analysis mode: ☐ By Context ☒ Association ☐ Sequences

Minimum Transaction Frequency to Support Associations:

☒ 5% of largest single item frequency ☐ Specify a count:

Maximum number of items in an association:

Minimum confidence for rule generation: %

Analysis mode

By Context	decide automatically by the model roles of the given data set
Association	Association discovery
Sequence	Sequence discovery

Minimum Transaction Frequency to Support Associations

- specifies a minimum level of support
- reduce the no. of rules considered

Maximum number of items in an association

- determines the maximum size of the item set to be considered
- including antecedent and consequent

Minimum confidence for rule generation

- specifies the minimum confidence level to generate a rule

Sequences Tab

- for sequence discovery only

Association

Selected Output | Notes

Data | Variables | General | Sequences | Time Constraints | Sort | Output

Minimum Transaction Frequency to Support Sequence:

☒ 2% of total transaction count

☐ Specify as a percentage: %

☐ Specify smallest count to use:

Number of items in longest chain:

Minimum Transaction Frequency to Support Sequence

- filter out sequences that occur too infrequently in the database

Number of items in longest chain

- set the maximum number of items to include in a sequence
- including antecedent and consequent

Time Constraints Tab

- for sequence discovery only

Association

Data | Variables | General | Processing Options | Sequences

Time Constraints | Sort | Output | Selected Output | Notes

Transaction Window Length:

☒ Maximum duration

☐ Specify duration to use:

Consolidate time differences < :

Transaction Window Length

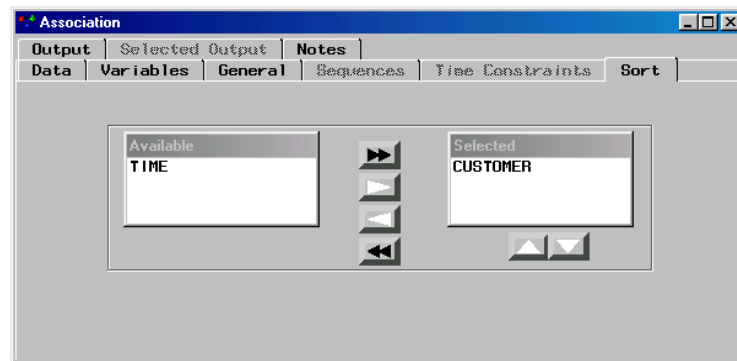
- max. length of time for the sequence

Consolidate time differences

- buying at 7:00 am, 12:00 pm and 9:00 pm → buying at the same day

Sort Tab

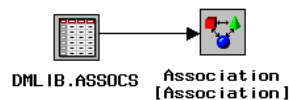
- set the primary and secondary sort keys for the ID variables (more than one ID variables)



- The **first** variable in the *Selected* list box is the **primary** sort key. All **other** ID variables are **secondary** (and higher order) sort keys.
- For example, consider that you have defined the following id variables as primary and secondary in the *Selected* list box:

Primary	Secondary	Objectives
STORE	CUSTOMER	explores the purchasing habits of customers within stores
CUSTOMER	STORE	explores store sales information for each customer

Association Discovery Example



Data Set:

SAMP5IO.ASSOCS

CUSTOMER → id

TIME → sequence

PRODUCT → target

Association : set *Analysis mode* (in *General* tab) to *Association*

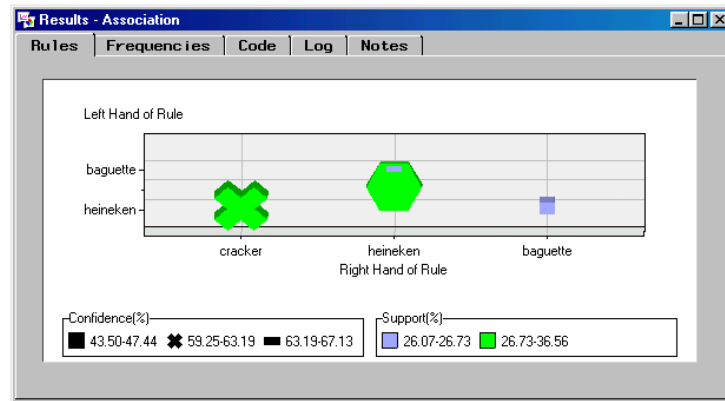
Interpreting the Results from an Association Discovery

Rules Tab

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	1.25	36.56	61.00	366.00	heineken ==> cracker
2	2	1.25	36.56	75.00	366.00	cracker ==> heineken
3	2	1.11	26.07	43.50	261.00	heineken ==> baguette
4	2	1.11	26.07	66.58	261.00	baguette ==> heineken
5	2	1.35	25.67	80.82	257.00	soda ==> heineken
6	2	1.35	25.67	42.83	257.00	heineken ==> soda
7	2	1.11	25.57	54.12	256.00	olives ==> heiring
8	2	1.11	25.57	52.67	256.00	heiring ==> olives
9	2	1.38	25.17	42.00	252.00	heineken ==> artichok
10	2	1.38	25.17	82.62	252.00	artichok ==> heineken
11	2	1.62	25.07	78.93	251.00	soda ==> cracker
12	2	1.62	25.07	51.43	251.00	cracker ==> soda
13	2	1.31	24.88	51.23	249.00	heiring ==> baguette

- *Table*
 - lists the relations (no. of items), lift, support, confidence and transaction count for each rule

- *Plot*
 - to display a plot of the rules, use the pull-down menu *View* → *Graph*
 - select first 3 rows from table, and then right click → *view* → *Graph*



Sequence Discovery Example

- set Analysis mode (in General tab) to Sequences

Results - Sequence

Rules | Frequencies | Code | Log | Notes

	Chain Length	Support(%)	Confidence(%)	Transaction Count	Rule
1	2	33.67	69.06	337	cracker ==> heineken
2	2	23.48	48.35	235	hering ==> heineken
3	2	23.28	49.26	233	olives ==> bourbon
4	2	22.88	47.12	229	hering ==> corned_b
5	2	22.58	46.50	226	hering ==> olives
6	2	22.48	57.40	225	baguette ==> heineken
7	2	21.98	69.18	220	soda ==> cracker
8	2	21.98	56.12	220	baguette ==> hering
9	2	21.98	46.51	220	olives ==> turkey
10	2	21.78	68.55	218	soda ==> heineken
11	2	21.68	73.31	217	coke ==> ice_crea
12	2	21.28	52.85	213	bourbon ==> cracker
13	2	20.98	53.71	210	corned_b ==> olives
14	2	20.88	53.32	209	baguette ==> avocado

- event chain is shown
- use *View* → *Subset Table...* to view subset rules
- e.g. view rules with *Chain Length* = 3
 1. select *Chain Length* from *Available Column*
 2. select *EQ operator*
 3. select *<LOOKUP distinct values>* → 3
 4. press *OK* button

