UNIVERSITÉ ANTONINE
Faculté d’ingénieurs en Informatique, Multimédia, Réseaux & Télécommunications

Data Mining – Implementation of Sequence Clustering

Matière : System décisionnel

Effectué par :	NOM Prénom	INF#	Option
	MATTA Elie	Privacy	OGL
	et al.	applied	

Introduction

This is a follow up report for our project: Sequence Clustering.

We would like to thank our teacher who gave us an extended week to complete our implementation. We based our essential work on the document of the "System decisionnel" which gave us a wider and a better idea about the right steps to complete our implementation.

Our project is about Sequence Clustering algorithm, to make it even easier to understand such an algorithm we decided to implement this algorithm in "Business Intelligence Development Studio" using the AdvanduteWorsDW database.

In the proceeding pages, we will guide you step-by-step into all the different parts and steps we ran into while implementing such algorithm using the upcoming screenshots.

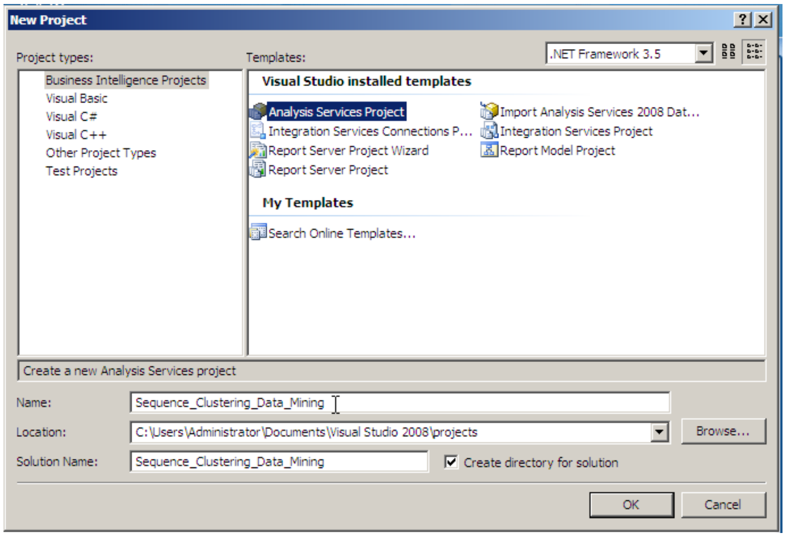
Note that more valuable information is included in the video "**sequence clustering implementation.avi**" included in the DVD, so don't hesitate to check the video recorded to follow the easy step-by-step documentation.

Implementation

I. Creating the Analysis Service Project

Step1:

- Open Business Intelligence Development Studio.
- Select New→ Project from the File menu.
- Select Analysis Services Project. We have named it “Sequence_Clustering_Data_Mining” as shown in Figure 1.
- Then click OK.


New Project
Project types:
Business Intelligence Projects
Visual Basic
Visual C#
Visual C++
Other Project Types
Test Projects
Templates:
.NET Framework 3.5
Visual Studio installed templates
Analysis Services Project
Integration Services Connections P...
Report Server Project Wizard
Report Server Project
Import Analysis Services 2008 Dat...
Integration Services Project
Report Model Project
My Templates
Search Online Templates...
Browse...
OK
Cancel

Create a new Analysis Services project

Name: Sequence_Clustering_Data_Mining

Location: C:\Users\Administrator\Documents\Visual Studio 2008\projects

Solution Name: Sequence_Clustering_Data_Mining

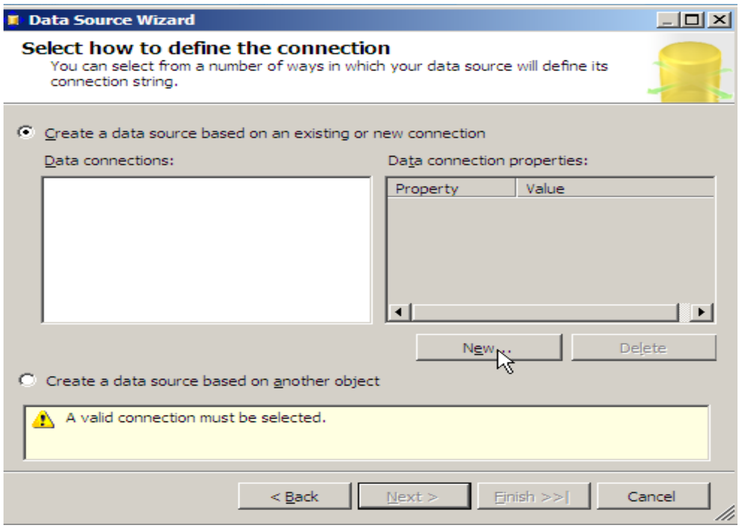
☑ Create directory for solution

Figure 1. New Analysis Services Project

II. Creating the Data Source

Step2:

- In the solution explorer, right click on “Data source” → new data source. “Data Source Wizard” dialog box opens.
- On the welcome page click Next.
- Click New to add a connection to the AdventureWorksDW database as shown in Figure 2.



Data Source Wizard

Select how to define the connection

You can select from a number of ways in which your data source will define its connection string.

● Create a data source based on an existing or new connection

Data connections:

Data connection properties:

Property	Value

○ Create a data source based on another object

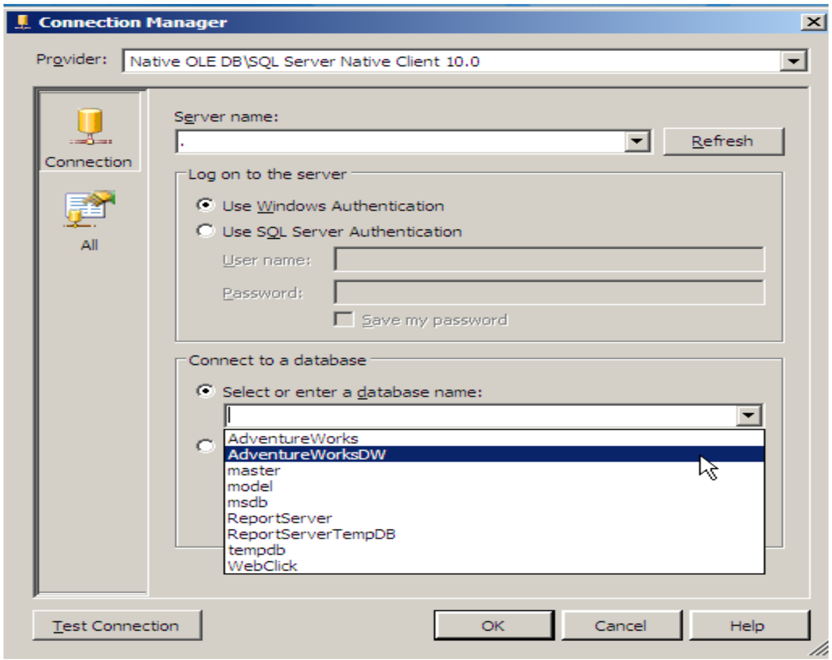
⚠ A valid connection must be selected.

Figure 2. New Data Source

Step3:

"Connection Manager" dialog box opens:

- In the Server name drop-down list, select the server where AdventureWorksDW is hosted (localhost="**.**")
- Use windows authentification.
- Select the database from the dropdown list → the **AdventureWorksDW** database
- Click OK to close the "Connection manager" dialog box.



Connection Manager

Provider: Native OLE DB\SQL Server Native Client 10.0

Server name: .

Log on to the server

● Use Windows Authentication

○ Use SQL Server Authentication

User name:

Password:

□ Save my password

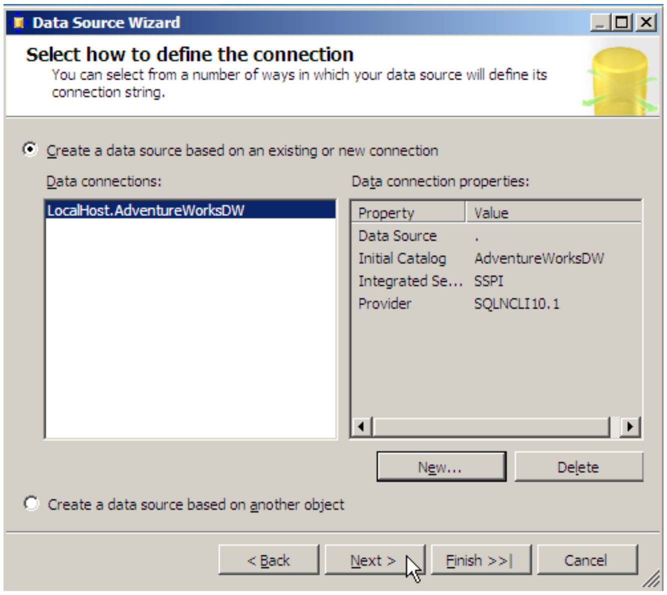
Connect to a database

● Select or enter a database name:

AdventureWorks
AdventureWorksDW
master
model
msdb
ReportServer
ReportServerTempDB
tempdb
WebClick

Figure 3. Connection Manager

Then you see the following figure then you click Next.



Data Source Wizard

Select how to define the connection

You can select from a number of ways in which your data source will define its connection string.

● Create a data source based on an existing or new connection

Data connections:

LocalHost.AdventureWorksDW

Data connection properties:

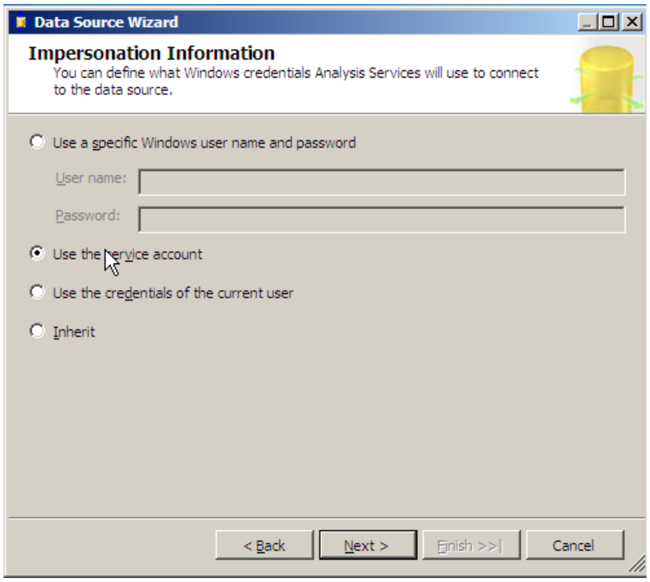
Property	Value
Data Source	.
Initial Catalog	AdventureWorksDW
Integrated Se...	SSPI
Provider	SQLNCLI10.1

○ Create a data source based on another object

Figure 4. Choosing the Data Connection

Step 4:

- Choose “Use the service account” combo box as shown in Figure 5.
- Click Next.



Data Source Wizard

Impersonation Information

You can define what Windows credentials Analysis Services will use to connect to the data source.

○ Use a specific Windows user name and password

User name:

Password:

● Use the service account

○ Use the credentials of the current user

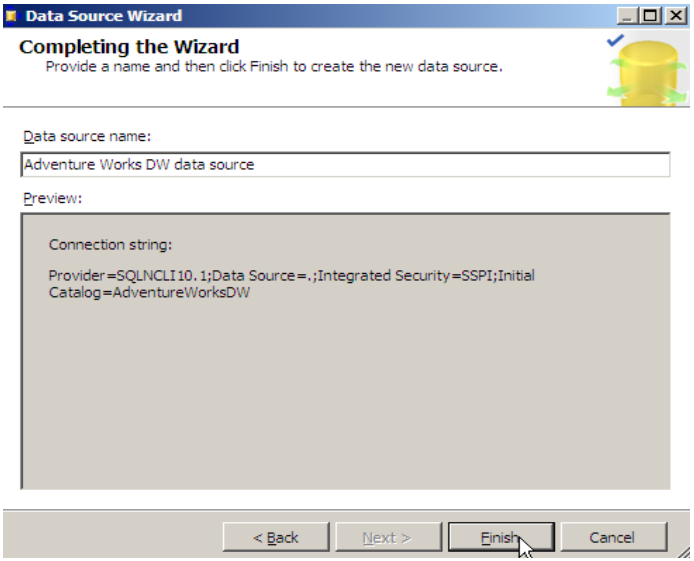
○ Inherit

Figure 5. using the service account.

Step 5:

Figure 6 shows this step.

- Give a name for the data source. We have named it “Adventure Works DW data source”.
- Click Finish.



Data Source Wizard

Completing the Wizard

Provide a name and then click Finish to create the new data source.

Data source name: Adventure Works DW data source

Preview:

Connection string:

```
Provider=SQLNCLI10.1;Data Source=.;Integrated Security=SSPI;Initial Catalog=AdventureWorksDW
```

Figure 6. Naming the data source

III. Creating Data Source View

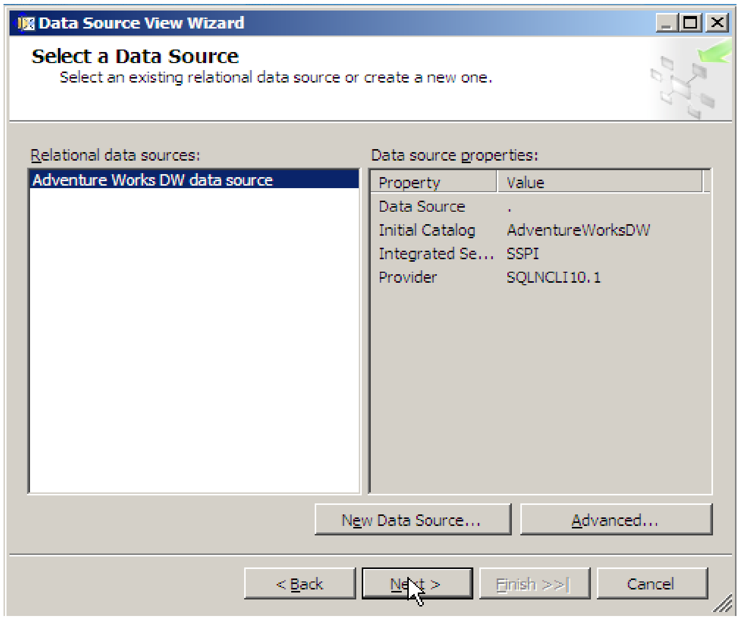
Step 6:

→See figure 7.

In the solution explorer, right click on “Data Source View”→ new data source view.

“Data Source View Wizard” dialog box opens.

- On the welcome page click Next.
- The “Adventure Works DW data source” data source created before is selected by default in the Relational data sources window.
- Click Next.



Select a Data Source

Select an existing relational data source or create a new one.

Relational data sources:

Adventure Works DW data source

Data source properties:

Property	Value
Data Source	.
Initial Catalog	AdventureWorksDW
Integrated Se...	SSPI
Provider	SQLNCLI10.1

Figure 7. Selecting the data source

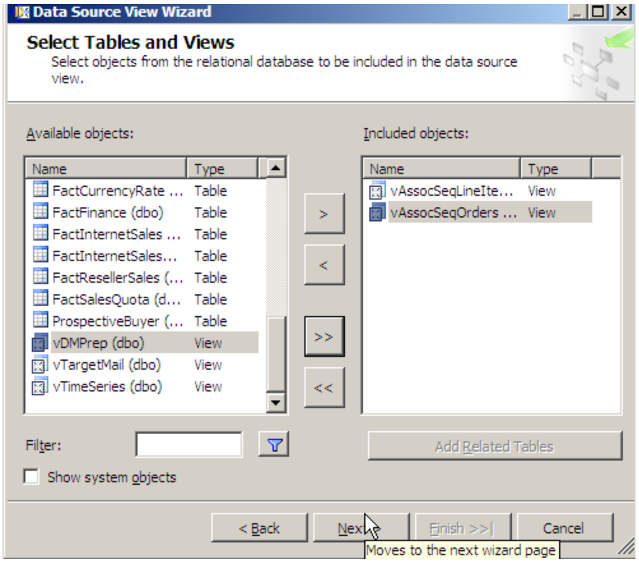
Step 7:

Here we choose what tables do we need for the model. So we have chosen:

- vAssocSeqOrders table and
- vAssocSeqLineItems table.

As shown in Figure 8.

- Then click next.



Select Tables and Views

Select objects from the relational database to be included in the data source view.

Available objects:

Name	Type
FactCurrencyRate ...	Table
FactFinance (dbo)	Table
FactInternetSales ...	Table
FactInternetSales...	Table
FactResellerSales (...	Table
FactSalesQuota (d...	Table
ProspectiveBuyer (...	Table
vDMPrep (dbo)	View
vTargetMail (dbo)	View
vTimeSeries (dbo)	View

Included objects:

Name	Type
vAssocSeqLineIte...	View
vAssocSeqOrders ...	View

Filter:

☐ Show system objects

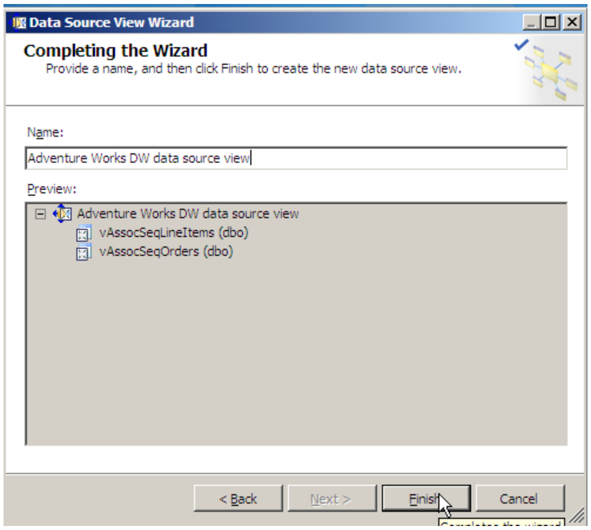
Moves to the next wizard page

Figure 8. Select tables and views

Step 8:

This step is shown in Figure 9.

- Give a name for the data source view. We have named it "Adventure Works DW data source view".
- Click Finish.



Completing the Wizard

Provide a name, and then click Finish to create the new data source view.

Name:
Adventure Works DW data source view

Preview:
- Adventure Works DW data source view
  - vAssocSeqLineItems (dbo)
  - vAssocSeqOrders (dbo)

Figure 9. Naming the data source view

So we have the following view shown in Figure 10. Do not forget to relate these two tables with a specific relationship (related by OrderNumber).

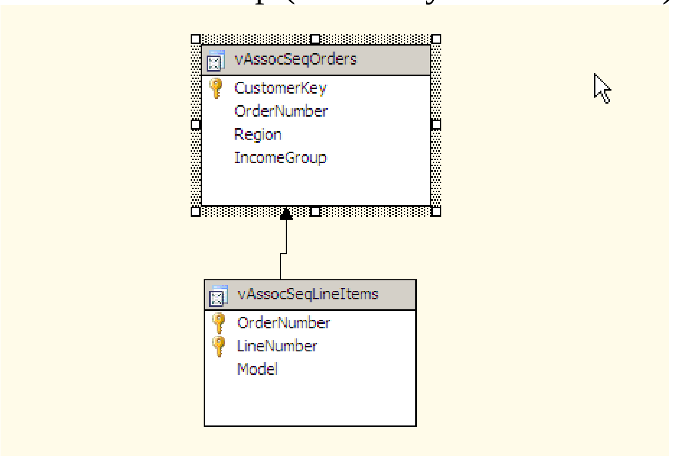


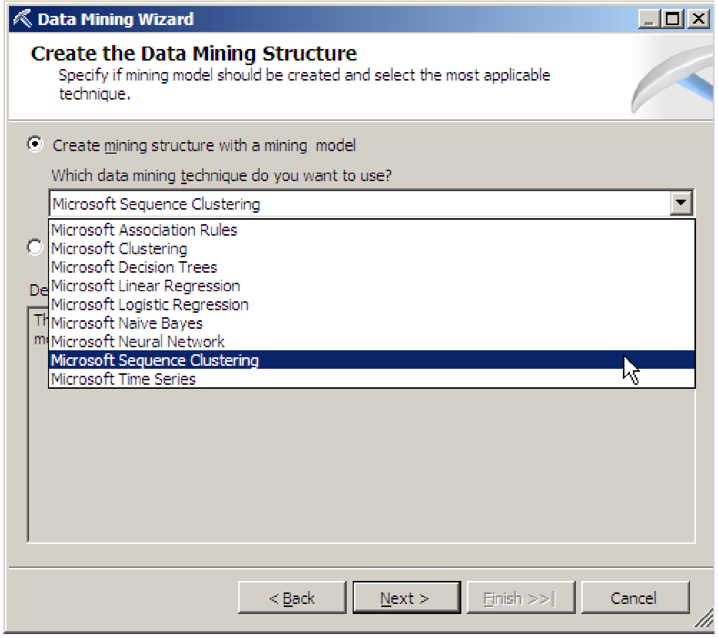
Figure 10. Adventure Works DW Data Source View

IV. Creating the mining structure

Step 9:

- In the solution explorer, right click “Mining Stucture” → new mining structure. “Data Mining Wizard” dialog box opens.
- On the welcome page click Next.
- Click “From existing relational database or data warehouse”.
- Click Next.

- Under "What data mining technique do you want to use?" click Microsoft Sequence Clustering as shown in Figure 11. Then click Next.



Data Mining Wizard

Create the Data Mining Structure

Specify if mining model should be created and select the most applicable technique.

● Create mining structure with a mining model

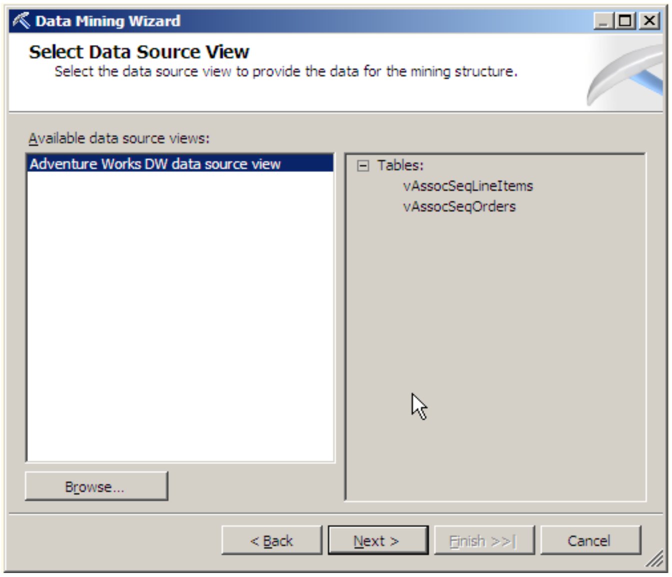
Which data mining technique do you want to use?

Microsoft Sequence Clustering

Figure 11. Choosing the data mining algorithm

<u>Step 10:</u> (see Figure 12)

- In this step we have to choose on which data source view we want to apply the data mining structure. So we choose the data source view created above the "Adventure Works DW data source view".
- Then click Next.



Data Mining Wizard

Select Data Source View

Select the data source view to provide the data for the mining structure.

Available data source views:

Adventure Works DW data source view

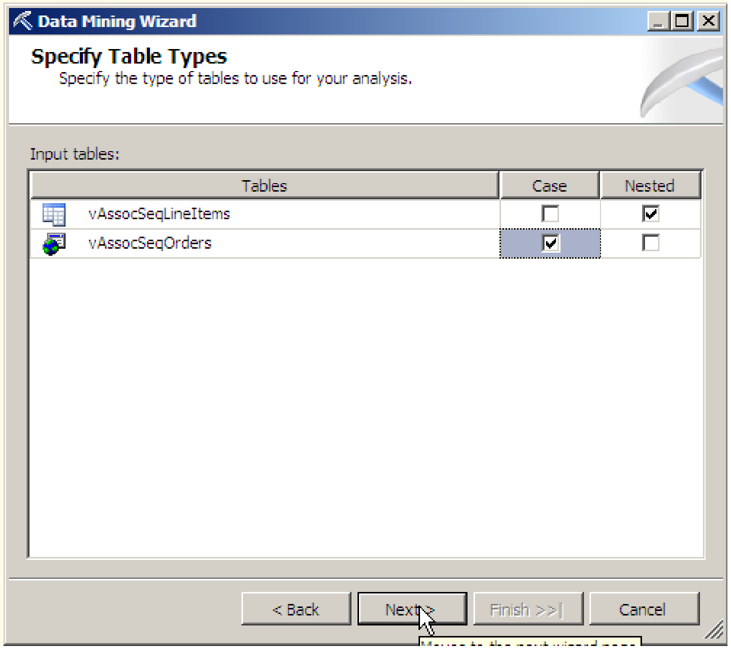
Tables:
vAssocSeqLineItems
vAssocSeqOrders

|**Figure 12.** Selecting Data Source View

Step 11:

See Figure 13.

- In this step we have to specify the tables type. Which is the nested and which are the cases. In our example vAssocSeqOrders is a case table and vAssocSeqLineItems is the nested table.
- Then click Next



Data Mining Wizard

Specify Table Types

Specify the type of tables to use for your analysis.

Input tables:

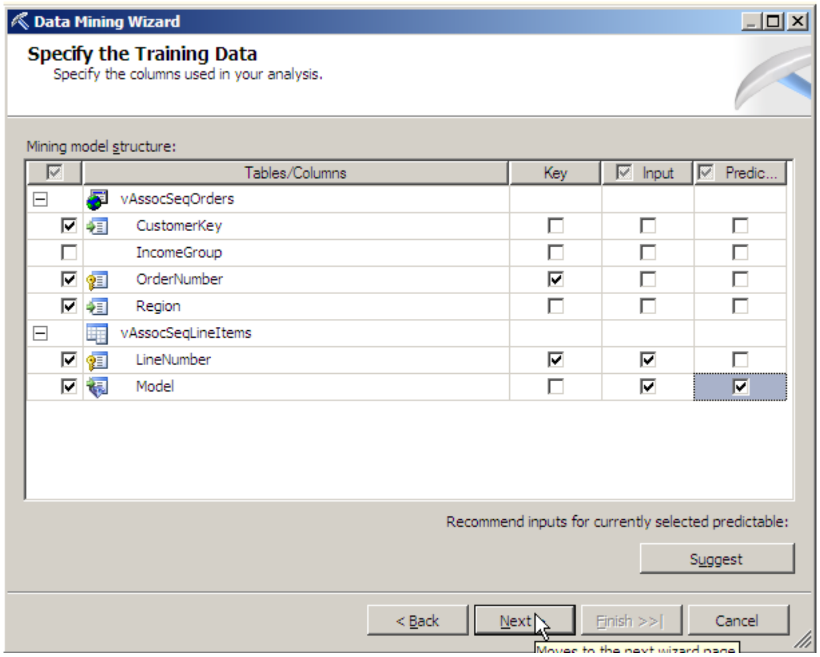
Tables	Case	Nested
vAssocSeqLineItems	☐	☑
vAssocSeqOrders	☑	☐

Figure 13. Specifying table types

Step 12:

In this step we can specify the columns used in our analysis.

- Clear the **key** check box next to **CustomerKey.**
- By default, OrderNumber and LineNumber are listed as Key types, which is correct.
- Select the **Input** and **Predictable** check boxes next to the **Model** columns. Make sure that the selection is the same as shown in Figure 14.
- Then click Next.



Data Mining Wizard

Specify the Training Data

Specify the columns used in your analysis.

Mining model structure:

☑	Tables/Columns	Key	☑ Input	☑ Predic...
⊟	vAssocSeqOrders			
☑	CustomerKey	☐	☐	☐
☐	IncomeGroup	☐	☐	☐
☑	OrderNumber	☑	☐	☐
☑	Region	☐	☐	☐
⊟	vAssocSeqLineItems			
☑	LineNumber	☑	☑	☐
☑	Model	☐	☑	☑

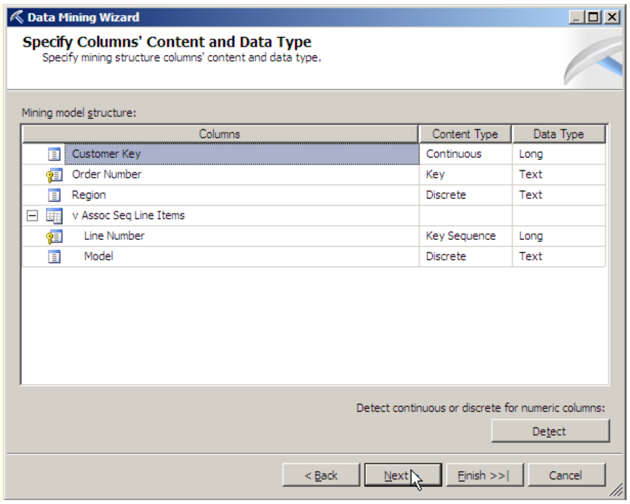
Recommend inputs for currently selected predictable:

Figure 14. Attribute specification for the sequence clustering mining structure.

Step 13:

See Figure 15.

- In this step we specify mining structure column's content and data type.
- Then click Next.



Data Mining Wizard

Specify Columns' Content and Data Type

Specify mining structure columns' content and data type.

Mining model structure:

Columns	Content Type	Data Type
Customer Key	Continuous	Long
Order Number	Key	Text
Region	Discrete	Text
⊟ v Assoc Seq Line Items		
Line Number	Key Sequence	Long
Model	Discrete	Text

Detect continuous or discrete for numeric columns:

Figure 15. Specifying column's content and data type

Step 14:

In this step we specify the number of cases to be reserved for model testing.
We determine the percentage of data for testing. By default it's 30%.
And also we determine the maximum number of cases in testing data set.
Then we click Next.

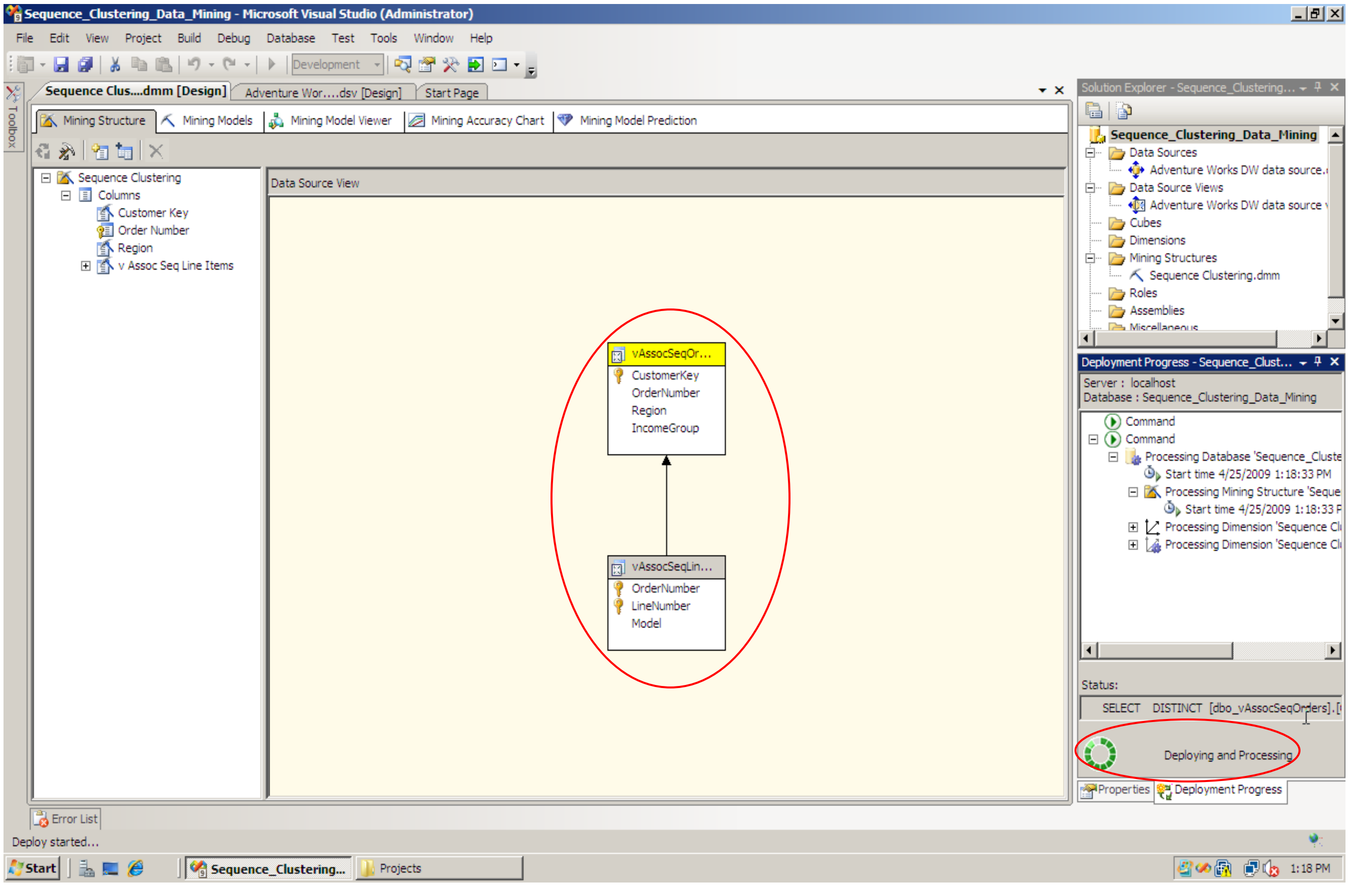
Step 15:

In this step we provide a name for the mining structure and the mining model.
We have named the both “Sequence Clustering”
Then we click Finish

Finally the new Sequence Clustering mining structure is displayed as shown in Figure 16.

Step 16:

- In the solution explorer, Right click on the project “Sequence_Clustering_Data_Mining” → Deploy



```
Server : localhost
Database : Sequence_Clustering_Data_Mining
Command
Command
Processing Database 'Sequence_Cluste
Start time 4/25/2009 1:18:33 PM
Processing Mining Structure 'Seque
Start time 4/25/2009 1:18:33 F
Processing Dimension 'Sequence Cl
Processing Dimension 'Sequence Cl
Status:
SELECT DISTINCT [dbo_vAssocSeqOrders].[
Deploying and Processing
Deploy started...
```

Figure 16. Sequence Clustering mining structure.

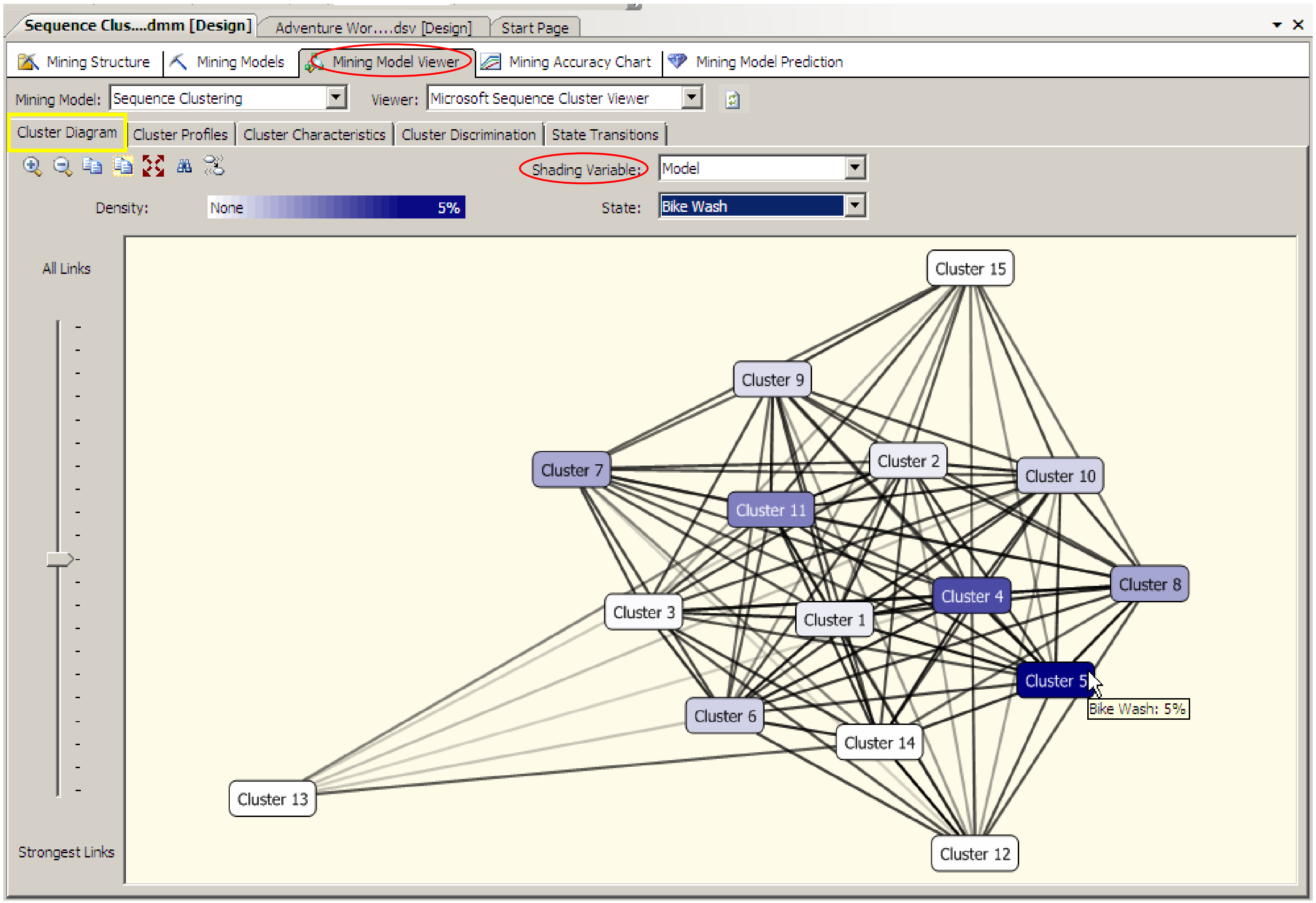
IV. Exploring mining model

We use the Sequence Clustering viewer to explore the mining model we created. To open the Sequence Clustering viewer, click Mining Model Viewer.

The Sequence Clustering viewer contains five tabs: Cluster Diagram, Cluster Profiles, Cluster Characteristics, Cluster Discrimination and State Transitions.

1) Cluster Diagram

The Cluster Diagram tab displays the clusters discovered by the algorithm in the database. The layout represents the cluster relationships. Similar clusters are grouped close together. By default, the node color represents the density of all cases in the cluster (the darker one contains the highest number of cases). We can also change the meaning of node color-coding so that it represents an attribute and state. For example, to generate the diagram shown in Figure 17, in the Shading Variable list, click Model, and in the State list, choose for example "Bike Wash". We can see that Cluster 5 contains the highest density of Bike Wash.



Mining Model: Sequence Clustering

Viewer: Microsoft Sequence Cluster Viewer

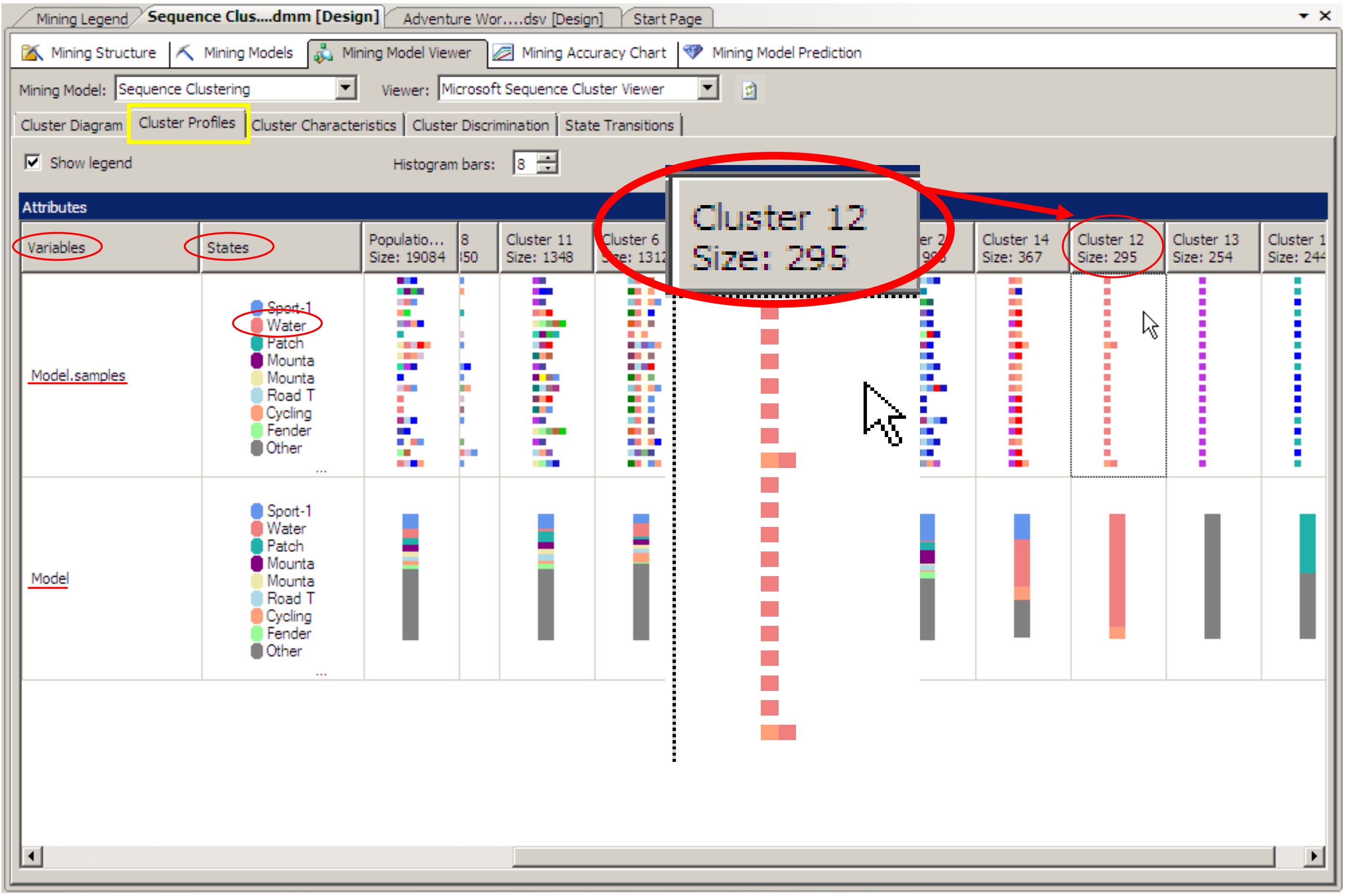
Shading Variable: Model

State: Bike Wash

Figure 17. Cluster Diagram tab of the Microsoft Sequence Clustering model

2) Cluster Profiles

The Cluster Profiles tab displays the sequences that exist in each cluster. The rows listed in the Variables column show the variable distributions for a cluster. In Figure 18, the Model.samples row represents sequence data, and the Model row describes the overall distribution of items in a cluster. Each line of the color sequences displayed in each cell of the Model.samples row represents the behavior of a randomly selected user in the cluster. Each color in the sequence histogram represents a product model.



Mining Model: Sequence Clustering

Viewer: Microsoft Sequence Cluster Viewer

☑ Show legend

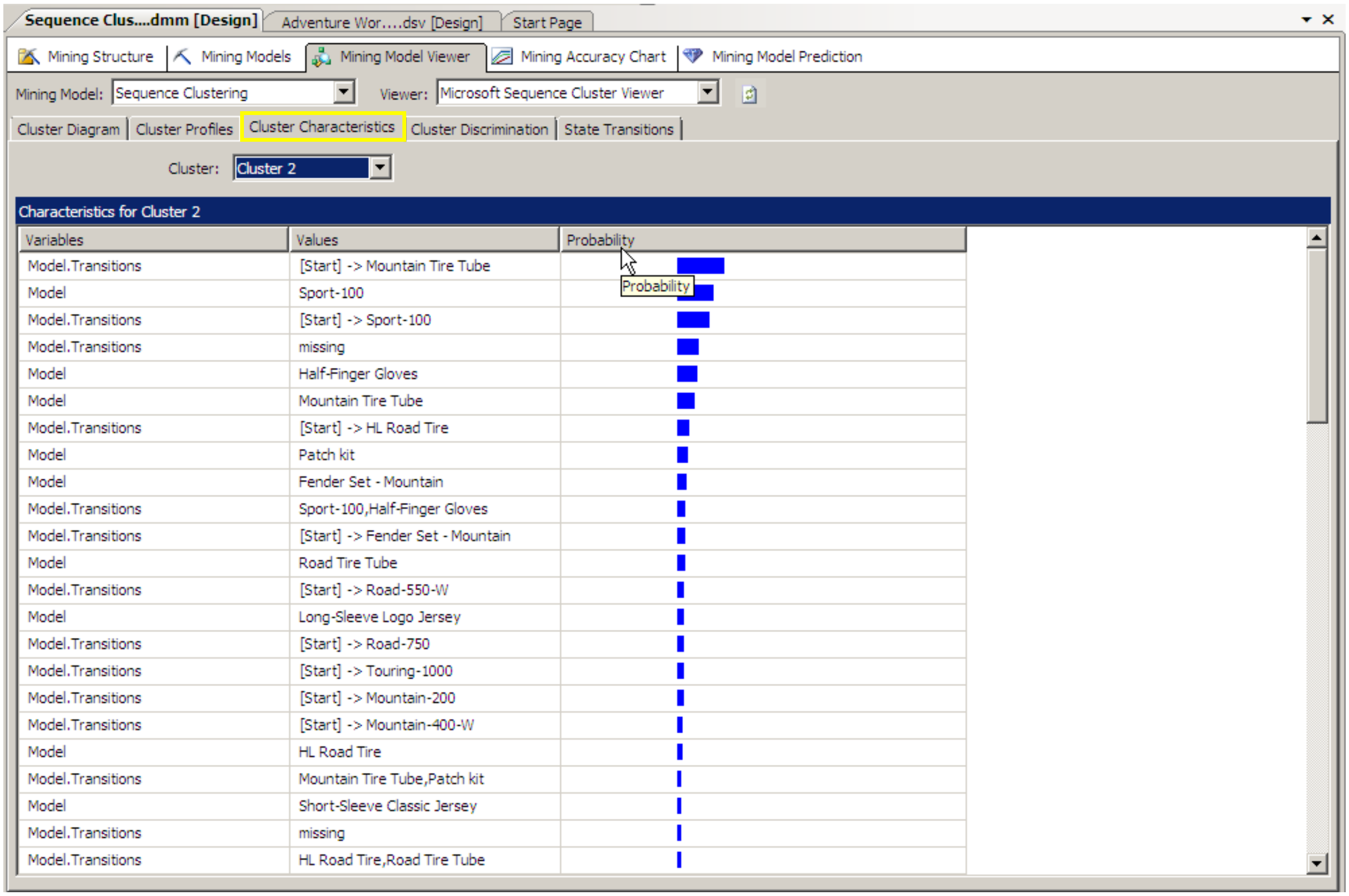
Histogram bars: 8

Figure 18. Cluster Profiles tab of the Microsoft Sequence Clustering model

For example, the pink color in cluster 12 represents the water bottle. The first color in most of the sequences is pink means that a customer is very likely to place water bottle in the shopping basket first.

3) Cluster Characteristics

The Cluster Characteristics tab summarizes the transitions between states in a cluster, with bars describing the importance of the attribute value for the selected cluster. For example, in Cluster 2, on of the most important profiles is that customers tend to place a Mountain tire tube in the shopping cart first.



Mining Model: Sequence Clustering

Viewer: Microsoft Sequence Cluster Viewer

Cluster: Cluster 2

Characteristics for Cluster 2

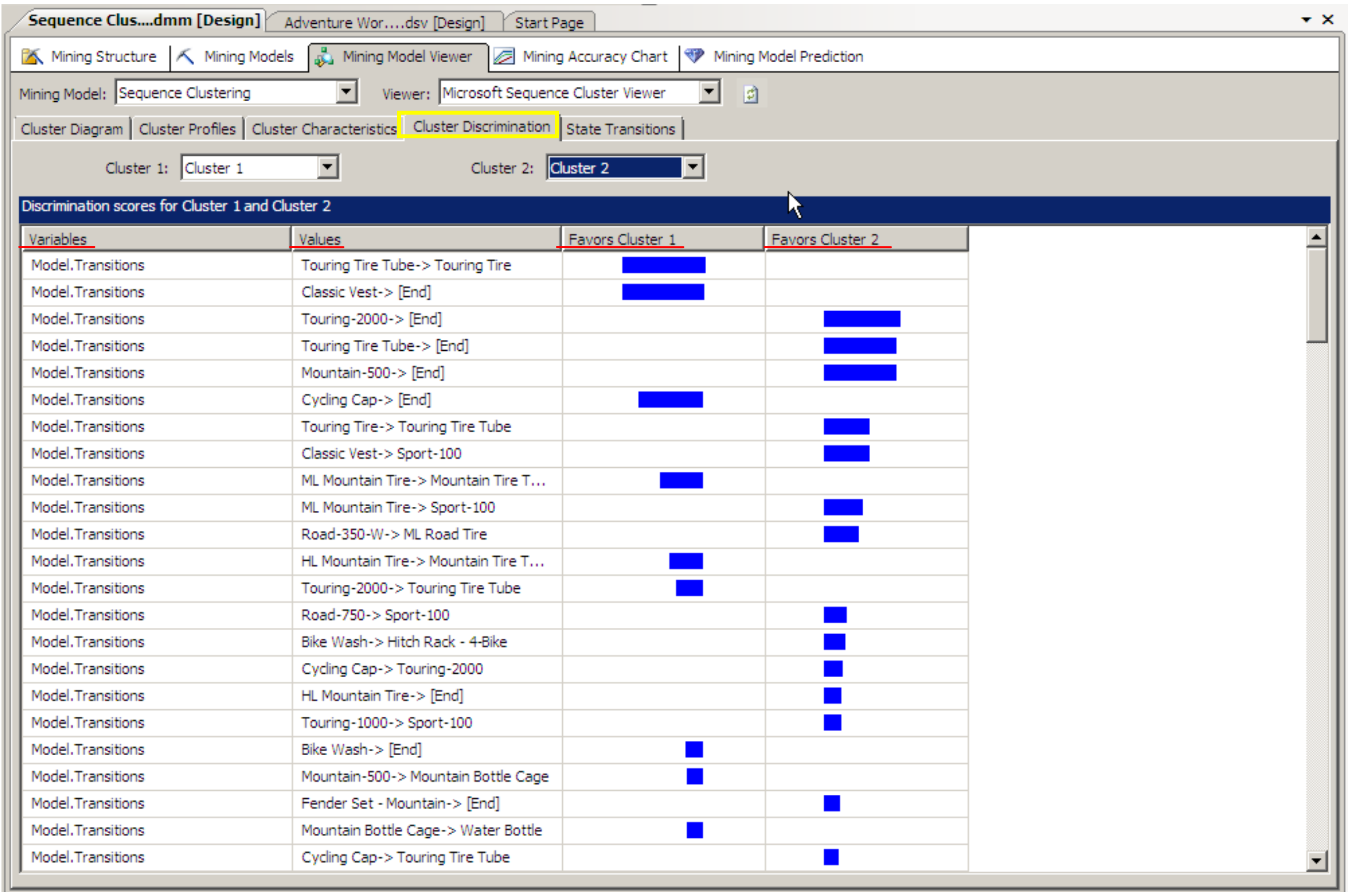
Variables	Values	Probability
Model.Transitions	[Start] -> Mountain Tire Tube	
Model	Sport-100	
Model.Transitions	[Start] -> Sport-100	
Model.Transitions	missing	
Model	Half-Finger Gloves	
Model	Mountain Tire Tube	
Model.Transitions	[Start] -> HL Road Tire	
Model	Patch kit	
Model	Fender Set - Mountain	
Model.Transitions	Sport-100,Half-Finger Gloves	
Model.Transitions	[Start] -> Fender Set - Mountain	
Model	Road Tire Tube	
Model.Transitions	[Start] -> Road-550-W	
Model	Long-Sleeve Logo Jersey	
Model.Transitions	[Start] -> Road-750	
Model.Transitions	[Start] -> Touring-1000	
Model.Transitions	[Start] -> Mountain-200	
Model.Transitions	[Start] -> Mountain-400-W	
Model	HL Road Tire	
Model.Transitions	Mountain Tire Tube,Patch kit	
Model	Short-Sleeve Classic Jersey	
Model.Transitions	missing	
Model.Transitions	HL Road Tire,Road Tire Tube	

Figure 19. Cluster Characteristics tab of the Microsoft Sequence Clustering model

4) Cluster discrimination

In the Cluster Discrimination tab, we can compare two clusters, determining which models favor which clusters. The tab contains four columns: Variables, Values, Favors Cluster (i), Favors Cluster (i). If the cluster favors a specific model, a blue bar appears in one of the Favors Cluster(i) columns, in the row of the model listed in the Variables column. The longer the blue bar, the more the model favors the cluster.

For example, Figure 20 compares Cluster 1 with Cluster 2. A customer who purchases a **Touring Tire Tube** is more likely to be in Cluster 2, and a customer who purchases a **Classic Vest** is more likely to be grouped into Cluster 1.



Mining Model: Sequence Clustering

Viewer: Microsoft Sequence Cluster Viewer

Cluster 1: Cluster 1

Cluster 2: Cluster 2

Discrimination scores for Cluster 1 and Cluster 2

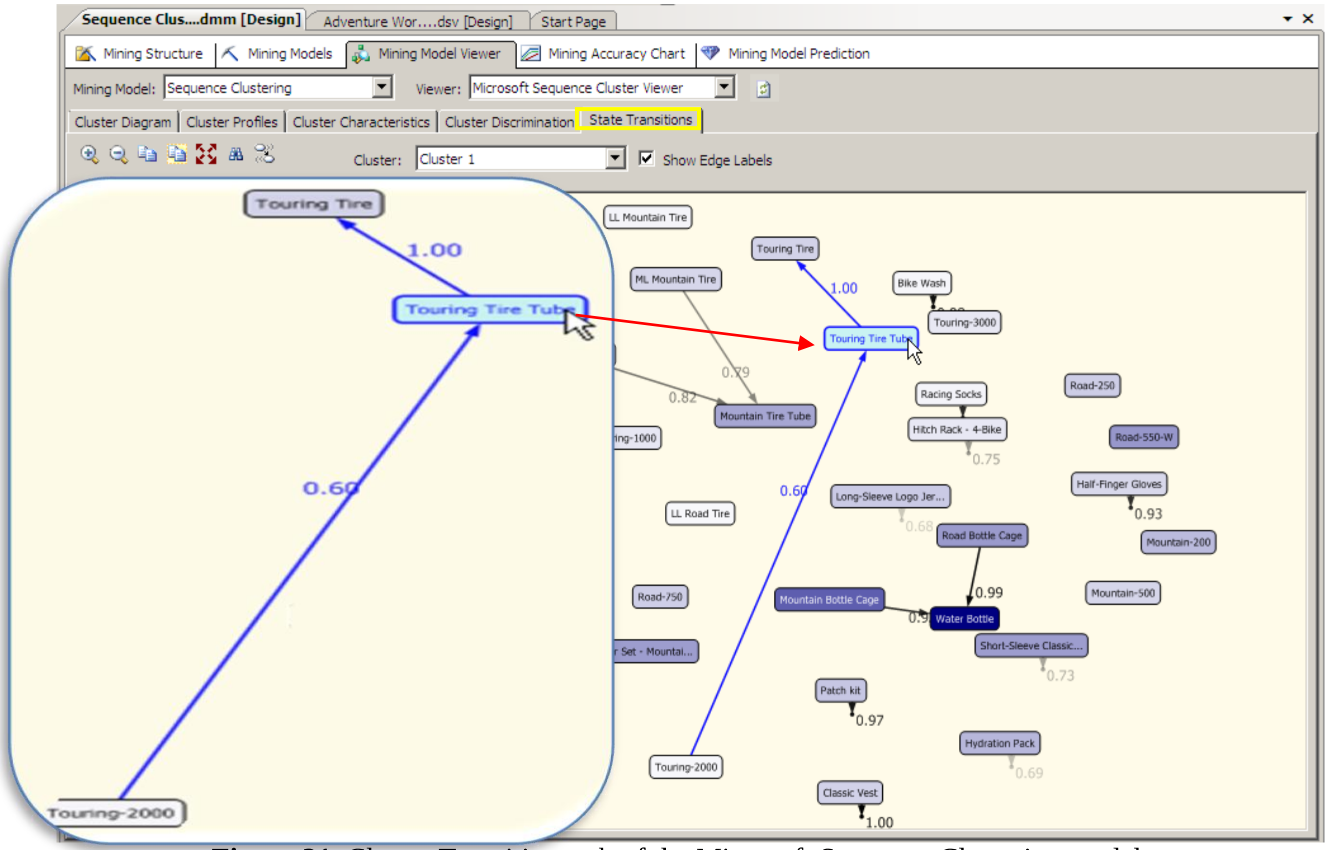
Variables	Values	Favors Cluster 1	Favors Cluster 2
Model.Transitions	Touring Tire Tube-> Touring Tire	■	
Model.Transitions	Classic Vest-> [End]	■	
Model.Transitions	Touring-2000-> [End]		■
Model.Transitions	Touring Tire Tube-> [End]		■
Model.Transitions	Mountain-500-> [End]		■
Model.Transitions	Cycling Cap-> [End]	■	
Model.Transitions	Touring Tire-> Touring Tire Tube		■
Model.Transitions	Classic Vest-> Sport-100		■
Model.Transitions	ML Mountain Tire-> Mountain Tire T...	■	
Model.Transitions	ML Mountain Tire-> Sport-100		■
Model.Transitions	Road-350-W-> ML Road Tire		■
Model.Transitions	HL Mountain Tire-> Mountain Tire T...	■	
Model.Transitions	Touring-2000-> Touring Tire Tube	■	
Model.Transitions	Road-750-> Sport-100		■
Model.Transitions	Bike Wash-> Hitch Rack - 4-Bike		■
Model.Transitions	Cycling Cap-> Touring-2000		■
Model.Transitions	HL Mountain Tire-> [End]		■
Model.Transitions	Touring-1000-> Sport-100		■
Model.Transitions	Bike Wash-> [End]	■	
Model.Transitions	Mountain-500-> Mountain Bottle Cage	■	
Model.Transitions	Fender Set - Mountain-> [End]		■
Model.Transitions	Mountain Bottle Cage-> Water Bottle	■	
Model.Transitions	Cycling Cap-> Touring Tire Tube		■

Figure 20. Cluster Discrimination tab of the Microsoft Sequence Clustering model

5) State transactions

On the State Transitions tab, we can select a cluster and browse through its state transitions. Each node represents a state of the model (such as Touring-2000). A line represents the transition between states, and each node is based on the probability of a transition. The background color represents the frequency of the node in the cluster.

For example, select **Cluster 2** from **Cluster.** As we can see in Figure 21, if users put a **Touring-2000** into his shopping cart, there is a probability of 0.60=60% (indicated by the blue arrow) that he will next put a **Touring Tire Tube** into the cart, and a probability of 1.00=100% that he will end his shopping by placing a **Touring Tire** into his shopping cart.



Mining Model: Sequence Clustering

Viewer: Microsoft Sequence Cluster Viewer

Cluster: Cluster 1

☑ Show Edge Labels

Figure 21. Cluster Transitions tab of the Microsoft Sequence Clustering model