Data Mining: Clustering

Applies to:
SAP BI 7.0. For more information, visit the [EDW homepage](http://www.sap.com)

Summary

This article deals with Data Mining and it explains the classification method ‘Clustering’ in detail. It also explains the steps for implementation of Clustering by creating a Model and an Analysis Process.

Author: Vishall Pradeep K.S
Company: Applexus Technologies (P) Ltd
Created on: 1 May 2011

Author Bio

Vishall Pradeep is working as SAP Technology Consultant with Applexus Technologies (P) Ltd. He has experience in SAP ABAP and SAP BI.
# Table of Contents

- Introduction .................................................................................................................. 3
- Clustering ....................................................................................................................... 3
- Creating a Model .......................................................................................................... 3
- Creating a Analysis Process for Training ..................................................................... 6
- Creating a Analysis Process for Executing a Prediction ............................................. 13
- Related Content .......................................................................................................... 21
- Disclaimer and Liability Notice .................................................................................... 22
Introduction

Data mining is to automatically determine significant patterns and hidden associations from large amounts of data. Data mining provides you with insights and correlations that had formerly gone unrecognized or been ignored because it had not been considered possible to analyze them. The data mining methods available in SAP BW allow you to create models according to your requirements and then use these models to draw information from your SAP BW data to assist your decision-making.

Clustering

Clustering allows you to segment data automatically into clusters. In a subordinate dataset, the system groups together associated data by forging formerly unknown links. This entails determining the criteria for clustering as well as the mappings between datasets.

Creating a Model

- Go to Transaction RSDMWB (Data Mining Workbench)

SAP Easy Access - User menu for APPLEXUS DEVELOPER

- Data Mining->Expand Clustering->Right Click Clustering->Create Model

Data Mining Workbench
- Choose the Model Name and Description
- The method name for which you are creating a model is displayed. You have three options for model field selection
  - To create the model fields manually, select the Manual option.
  - If you want to create a model that is similar to an existing model created previously, you can copy it choosing the Use Model as Template option. You can make minor changes to the copied version manually to suit your requirements.
  - To create a model from a query, choose Model Field Selection and select the query which you want use as a source for model fields. The InfoObjects contained in the selected query are available as model fields.

Queries and Views

Available Objects
- IC for Ekpo
- Purchase Order Test - Vishall
- Purchase Order Test - Vishall

Create Model

Model Name: ZPDT_MODEL_CLUST
Description: Model for Clustering
Method:
Options:
- Manual
- Use Model as Template
- Model Field Selection
  - BVW Query: I_IC_VEKO/REP_20110420063408
The screen shows the list of Fields and we can select and exclude fields in it.

In the Fields tab, to specify which characteristic is to be considered with which attributes and field parameters are used to specify weightings for the individual attributes. The system then establishes formerly unknown associations between the attribute values.
The Content types valid for a model field are dependent on the method that you are creating the model for and on the data type of the model field. The value type specified for a model field determines which entries can be made as Field Parameters and Field Values.

In the Parameters tab, we can specify how many clusters the system should create during training and by specifying conditions for interrupting the segmentation; you enhance the quality and performance of the segmentation.

Excluding the what if and export as PMML boxes, the fields list comprises the following:

- Model Parameters
  - No. of Clusters
  - Max. Distinct Values allowed

- Stopping Criteria
  - Max. no. of Iterations
  - Min. Fraction of Inter Cluster hops

Save and Activate the Model (we can only train orvaluate a model or use it for the prediction if the model has been activated.)

Creating a Analysis Process for Training

- Go to Transaction RSANWB (Analysis Process Designer)
• Choose General->Right Click->Create

![Analysis Process Designer](image1)

• Give the description to the APD

![Analysis Process Designer - Create -](image2)

• From the Data Sources, drag and drop the Query to the work area

![Analysis Process Designer - Create -](image3)
- It asks for a Popup and click on Choose Query

- From the Help, Select the query

- And Click “OK”
- The Query which is the data Source is added as below

- For the data target, drag the icon for the relevant data mining method in the work area

- Double click on data mining node to make the settings in the dialog box that appears
- Choose the required model from F4 Help
- Click on “ENTER” and Connect the two nodes

- To make an explicit field assignment, double click on the data flow arrow that connects the nodes
- Click on Automatic Assignment and choose Same Infoobject

---

<table>
<thead>
<tr>
<th>Method</th>
<th>Source Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same InfoObject</td>
<td></td>
</tr>
<tr>
<td>Same Data Element</td>
<td></td>
</tr>
<tr>
<td>Purchase order quantity (R/3)</td>
<td></td>
</tr>
<tr>
<td>Purchasing document number</td>
<td></td>
</tr>
<tr>
<td>Item number of purchasing document</td>
<td></td>
</tr>
</tbody>
</table>
• Click on Continue and Save and activate the APD
• While saving it will ask for a Technical Name

![Enter a technical name](image)

Analysis Process: ZPDT_MOD_CLUS

![Log Editor](image)

Display logs

- Execute analysis process ZPDT_MOD_CLUS, version A
- Source data is being read and the result is being calculated
- Too many values! Discrete model field 001_EBLEN ignored
- Data mining model successfully trained
- Data successfully written to the data target of the analysis process
- Processing completed successfully for analysis process ZPDT_MOD_CLUS

• Execute the APD
• The data is written to the data target and a log is displayed

![Display logs](image)

• To view the training results, in the context menu of data target, choose Data Mining Model → View Model Results

![Data Mining Model](image)
The Results will be shown as below

- By Using Value Distribution button you can "View Attribute Value Distribution as Graphic" and Attribute Data to "View Attribute Value Distribution as Tables" and Attribute Charts to "View Attribute Value Distribution across Clusters".
- The data mining model acquires the status Trained

Display Model ZPDT_MODEL_CLUST
Creating a Analysis Process for Executing a Prediction

- A model that you trained using historic data from a source can now be applied to a different set of data and the predicted output could be the best of three clusters
- Goto Transaction RSANWB (Analysis Process Designer)

Choose General->Right Click->Create

- Give the description to the APD
- From the Data Sources, drag and drop the Query to the work area

- It asks for a Popup and click on Choose Query

- From the Help, Select the query
- Drag the relevant prediction icon, that is, source for transformation, in the work area.

- Connect the two nodes.
- Double click on data mining node to make the settings in the dialog box that appears.
- Choose the required model from F4 Help.

- In Prediction Input Tab, Map the Available source field and Input field for prediction.
• In select Prediction Output Fields “Select the Fields”

![Image of Prediction Transformation]

• For the data target, drag the icon for the relevant data mining method in the work area (In this case I am downloading the predicted values to another Cluster Model)

• Create a cluster model in “RSDMWB” and Keep it in New State (Don’t Train the Model)

![Image of Analysis Process Designer - Create - DataMining: Prediction - Clustering]

• Double click the Target and Choose the model

![Image of Training a Data Mining Model]
• Connect the nodes
• Double Click on the Node Connection and assign the source and target structure fields

![Change Field Assignment](image)

Field Assignment:

Use the possible entries function in the right column to:
assign the source structure fields to the target structure fields.

- Same InfoObject
- Same Data Element

<table>
<thead>
<tr>
<th>Method</th>
<th>Source Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purchase order quantity (QTY)</td>
<td></td>
</tr>
<tr>
<td>Material</td>
<td></td>
</tr>
<tr>
<td>Purchasing document number</td>
<td></td>
</tr>
<tr>
<td>Item number of purchasing doc</td>
<td></td>
</tr>
</tbody>
</table>

• Save and activate the APD
• While saving it will ask for a Technical Name

![Enter a technical name](image)

Enter technical name: ZDM_CLS_PREDICT

• Execute the APD

![Display logs](image)

- Execute analysis process ZDM_CLS_PREDICT, version A
- Source data is being read and the result is being calculated
- Data mining model successfully trained
- Data successfully written to the data target of the analysis process
- Processing started for analysis process ZDM_CLS_PREDICT
- Too few values present. Discrete model field 0BASE_UOM ignored
- Too many values! Discrete model field 001_EBELN ignored
- Processing completed successfully for analysis process ZDM_CLS_PREDICT
To view the Prediction results, in the context menu of data target, choose Data Mining Model → View Model Results

The Results will be shown as below

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 001</td>
<td>62</td>
</tr>
<tr>
<td>Cluster 002</td>
<td>4</td>
</tr>
<tr>
<td>Cluster 003</td>
<td>56</td>
</tr>
<tr>
<td>Cluster 004</td>
<td>27</td>
</tr>
<tr>
<td>Cluster 005</td>
<td>10</td>
</tr>
<tr>
<td>Overall</td>
<td>160</td>
</tr>
</tbody>
</table>

Influence Chart
- Click on Distribution button you can "View Attribute Value Distribution as Graphic"

- Click on Attribute Data to "View Attribute Value Distribution as Tables"

### Purchase order quantity (SAP D)

<table>
<thead>
<tr>
<th></th>
<th>Cluster 001</th>
<th>Cluster 002</th>
<th>Cluster 003</th>
<th>Cluster 004</th>
<th>Cluster 005</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00 TO 442.00</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>442.00 TO 884.00</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>29</td>
</tr>
<tr>
<td>884.00 TO 1326.00</td>
<td>30</td>
<td>0</td>
<td>48</td>
<td>16</td>
<td>0</td>
<td>94</td>
</tr>
<tr>
<td>1326.00 TO 1768.00</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1768.00 TO 3973.00</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3973.00 TO 4424.00</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### Material

<table>
<thead>
<tr>
<th></th>
<th>Cluster 001</th>
<th>Cluster 002</th>
<th>Cluster 003</th>
<th>Cluster 004</th>
<th>Cluster 005</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>100-131</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>100-139</td>
<td>62</td>
<td>4</td>
<td>56</td>
<td>27</td>
<td>10</td>
<td>159</td>
</tr>
</tbody>
</table>

### Item number of purchasing docu

<table>
<thead>
<tr>
<th></th>
<th>Cluster 001</th>
<th>Cluster 002</th>
<th>Cluster 003</th>
<th>Cluster 004</th>
<th>Cluster 005</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00 TO 14.00</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>19.00 TO 24.00</td>
<td>52</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>55</td>
</tr>
<tr>
<td>29.00 TO 34.00</td>
<td>0</td>
<td>0</td>
<td>56</td>
<td>25</td>
<td>0</td>
<td>52</td>
</tr>
<tr>
<td>54.00 TO 60.00</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
• Click on Attribute Charts to "View Attribute Value Distribution across Clusters"

• Click on Intra Cluster Attribute Value Distribution Index to view attribute index

• Click on Inter and Intra Cluster Distance Information
Related Content

Clustering

Data Mining

Data Mining-Decision Trees

For more information, visit the EDW homepage
Disclaimer and Liability Notice

This document may discuss sample coding or other information that does not include SAP official interfaces and therefore is not supported by SAP. Changes made based on this information are not supported and can be overwritten during an upgrade. SAP will not be held liable for any damages caused by using or misusing the information, code or methods suggested in this document, and anyone using these methods does so at his/her own risk.

SAP offers no guarantees and assumes no responsibility or liability of any type with respect to the content of this technical article or code sample, including any liability resulting from incompatibility between the content within this document and the materials and services offered by SAP. You agree that you will not hold, or seek to hold, SAP responsible or liable with respect to the content of this document.