



Data Mining – LAB 3

Clustering – K-Means

Data File

- In Virtualbox image:
 - c:\data\data-Lab3-lab3.csv
 - c:\data\data-Lab3-Link.xls
 - c:\data\data-Lab3-Shop_Info.xls
- Download from:
 - www.comp.polyu.edu.hk/~csamak/data/Lab3.zip
 - www.comp.polyu.edu.hk/~csamak/data/data-Lab3.mdb

For *data-Lab3.mdb*, it does not work probably in the **VM**. So, do the lab with the “Lab3.zip” if you use the virtual machine.

Import Data from MS Excel File

- Use 3 “Var. File” nodes to import the three data files

The image displays a data import interface with three file nodes on the left and their corresponding configuration windows on the right.

File Nodes:

- data-Lab3-lab3.csv
- data-Lab3-Link.xls
- data-Lab3-Shop_Info...

Configuration Windows:

data-Lab3-lab3.csv

File: C:\Data\data-Lab3-lab3.csv

Field	Type
TID	Typeless
dt	Range
gp1	Flag
gp2	Range
ref_no	Range
cl	Range
prod_cd	Set

data-Lab3-Link.xls

File: C:\Data\data-Lab3-Link.xls

Field	Type	Values	Missing	Check	Direction
TID	Typeless			None	None
SHOP_...	Set	A,B,C,D,...		None	In

data-Lab3-Shop_Info...

File: C:\Data\data-Lab3-Shop_Info...

Field	Type	Values	Missing	Check	Direction
dist_cd	Set	A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,...		None	In
shop_cd	Set	A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,...		None	In
staffs	Range	[5.0,300.0]		None	In
manager	Range	[1.0,3.0]		None	In
area	Set	"250","270","280","3...		None	In

Import Data

- Merge the three tables into ONE

The screenshot shows a data management application window titled "Table (12 fields, 1,241 records)". The window contains a table with the following columns: SHOP_CD, TID, dt, gp1, gp2, ref_no, cl, prod_cd, dist_cd, staffs, manager, and area. The first 10 rows of data are visible.

	SHOP_CD	TID	dt	gp1	gp2	ref_no	cl	prod_cd	dist_cd	staffs	manager	area
1	A	CEA/1/367/02	2002-09-04 10:30:00	N	2	210	109 P30	CHG	10.0...	1.000	300	
2	A	CDC/11/166/02	2002-03-14 15:10:00	N	2	206	109 P30	CHG	10.0...	1.000	300	
3	A					206	109 P34	CHG	10.0...	1.000	300	
4	A					766	109 P27	CHG	10.0...	1.000	300	
5	A					210	109 P27	CHG	10.0...	1.000	300	
6	A					767	109 P7	CHG	10.0...	1.000	300	
7	A					210	109 P7	CHG	10.0...	1.000	300	
8	A					766	109 P28	CHG	10.0...	1.000	300	
9	A					210	109 P28	CHG	10.0...	1.000	300	
10	A					767	109 P30	CHG	10.0...	1.000	300	

An inset diagram illustrates the process of merging three tables into one. It shows three input tables on the left, each with a specific number of fields: "Table" (6 Fields), "Table" (1 Fields), and "Table" (5 Fields). Arrows from these tables point to a central "Merge" icon, which then points to a final "Merge" icon, representing the output of the merge operation.

To merge tables, the common key should be defined first. Can you form the table as shown?

Data Understanding/Cleaning

- What are the errors in the data set?
- Suggest methods to correct them.

Data Transformation

Table (14 fields, 1,241 records)

File Edit Generate

New columns added

	dt	gp1	gp2	ref_no	cl	prod_cd	dist_cd	staffs	manager	area	Weekday	Time
1	2002-09-04 10:30:00	N	2	210	109	P30	CHG	10.0...	1.000	300	Wednesday	10
2	2002-03-14 15:10:00	N	2	206	109	P30	CHG	10.0...	1.000	300	Thursday	15
3	2002-08-05 09:50:00	N	2	206	109	P34	CHG	10.0...	1.000	300	Monday	9
4											Thursday	13
5											Thursday	13
6											Thursday	13
7											Thursday	13
8											Thursday	12
9											Thursday	12
10											Saturday	12
11											Saturday	12
12	2002-09-19 19:20:00	N	2	207	109	P27	CHG	10.0...	1.000	300	Thursday	19
13	2002-09-19 18:51:00	N	2	766	109	P30	CHG	10.0...	1.000	300	Thursday	18
14	2002-09-19 18:51:00	N	2	210	109	P30	CHG	10.0...	1.000	300	Thursday	18
15	2002-12-19 07:40:00	N	2	206	109	P34	CHG	10.0...	1.000	300	Thursday	7
16	2002-12-19 07:40:00	N	2	206	109	P27	CHG	10.0...	1.000	300	Thursday	7
17	2002-03-08 07:45:00	N	2	206	109	P30	CHG	10.0...	1.000	300	Friday	7
18	2002-09-17 14:50:00	N	2	213	109	P30	CHG	10.0...	1.000	300	Tuesday	14
19	2002-09-17 14:50:00	N	2	210	109	P30	CHG	10.0...	1.000	300	Tuesday	14
20	2002-09-17 14:20:00	N	2	213	109	P30	CHG	10.0...	1.000	300	Tuesday	14

Diagram illustrating data transformation steps:

```

graph LR
    Merge --> Weekday
    Weekday --> Time
  
```

Weekday: `datetime_day_name(datetime_weekday(dt))`
 Time: `datetime_hour(datetime_time(dt))`

OK

Data Transformation – Discretization

- Divide “Time” into 3 intervals (binning) and rename the values with node, Reclassify

The image shows two Alteryx tool windows: **Binning** and **Reclassify**.

Binning Tool Configuration:

- Bin fields:** Time
- Binning method:** Fixed-width
- Fixed-width Binning:**
 - Name extension:** _BIN
 - Add as:** Suffix
 - Bin width:** 10.0
 - No. of bins:** 3
- Bin thresholds:** Always recompute

Reclassify Tool Configuration:

- Mode:** Single
- Reclassify into:** Existing field
- Reclassify field:** Time_BIN
- New field name:** Reclassify1
- Reclassify values:**

Original value	New value
1	Morning
2	Afternoon
3	Evening

For unspecified values use: Original... Default... undef

Data Transformation – Discretization

- Divide “staffs” into 5 intervals (Fixed-width)

Table (16 fields, 1,241 records)

	prod_cd	dist_cd	staffs	manager	area	Weekday	Time	Time_BIN	staffs_BIN
524	P30	CSW	50.0...	2.000	400	Tuesday	7 Morning		1
525	P39	CSW	50.0...	2.000	400	Friday	10 Afternoon		1
526	P27	CSW	50.0...	2.000	400	Monday	16 Evening		1
527	P39	CSW	50.0...	2.000	400	Thursday	16 Evening		1
531	P2	CSW	50.0...	2.000	400	Thursday	16 Evening		1
532	P27	CSW	50.0...	2.000	400	Monday	16 Evening		1
533	P39	CWN	80.0...	2.000	490	Friday	12 Afternoon		2
534	P27	CWN	80.0...	2.000	490	Friday	12 Afternoon		2
535	P34	CWN	80.0...	2.000	490	Saturday	10 Afternoon		2
536	P27	CWN	80.0...	2.000	490	Saturday	10 Afternoon		2
537	P24	CWN	80.0...	2.000	490	Saturday	10 Afternoon		2
538	P17	CWN	80.0...	2.000	490	Saturday	10 Afternoon		2
539	P12	CWN	80.0...	2.000	490	Saturday	10 Afternoon		2
540	P44	CWN	80.0...	2.000	490	Saturday	10 Afternoon		2
541	P30	CWN	80.0...	2.000	490	Sunday	7 Morning		2
542	P30	CWN	80.0...	2.000	490	Sunday	8 Afternoon		2
543	P39	CWN	80.0...	2.000	490	Saturday	16 Evening		2

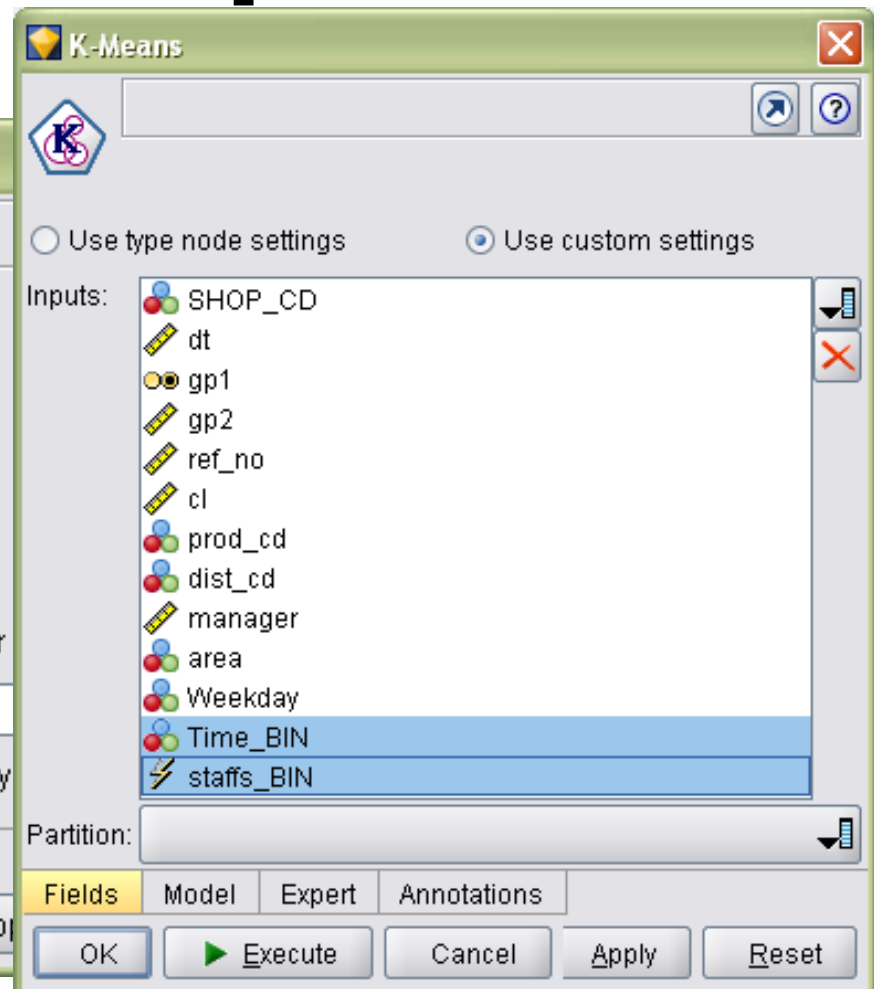
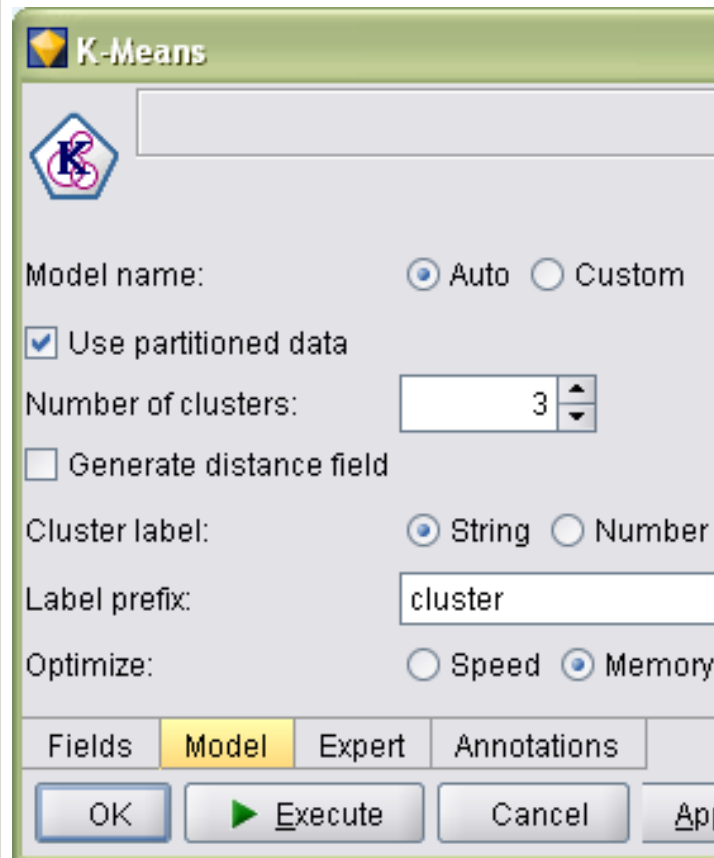
Finally, you may get the dataset as shown.

Table Annotations

OK

Clustering – Divide “Customers into 3 Clusters

- Add node, [**K-Means**] and set number of cluster to [**3**]



Result

