

MIS772

Predictive Analytics

Workshop: Data Clustering

k-Means clustering, optimisation of clustering, cluster visualisation
with PCA



Workshop Plan

Objectives:

Your task is to create a cluster model of people coming to a shopping mall, based on information collected via a survey. While we will utilise a k-means cluster model, with some adaptations the tasks are applicable to other clustering methods as well, e.g. k-medoids.

Data Set:

Marketing-Keel.zip

Use file “marketing.csv”

Original Data from KEEL:

<https://sci2s.ugr.es/keel/dataset.php?cod=163>

Method:

Attend the workshop, follow the tutor’s demo and instructions, take notes. Note that the class and online seminar will be recorded and their videos linked to the CloudDeakin topic for later access and study.

1 Acquire data for clustering

- (a) Load the data and unzip
- (b) Read and explore the data set, and store

2 Create a k-means clustering model

- (a) Select all attributes
- (b) Normalise and replace missing values
- (c) Add k-means with default parameters
- (d) Daisy chain cluster performance operators
- (e) Experiment with different number of clusters k , save

3 Optimise the cluster model

- (a) Adapt the previous process for cluster optimisation
- (b) Use grid optimisation to monitor k-means k
- (c) Log all performance criteria while changing k
- (d) Plot performance and find optimum k , save

4 Further analysis

- (a) Use the first process
- (b) Enter the optimum k into k-means
- (c) Add cluster visualiser
- (d) Run, interpret cluster visualiser results, save
- (e) Discuss your insights in class
- (f) *Challenge: Use PCA to plot and diagnose clusters*

K-Means Clustering

First, we will create a process responsible for data preparation and clustering. Use k-Means with defaults. Daisy-chain two performance operators to find *Davis-Bouldin* (DB) and *Within Sum of Squares* (WSS) measures.

Set k-means k. Run the process. Observe results and the clustering performance. Save.

id	cluster	Sex	Marit...	Age	Educat
1	cluster_0	1	0	0.667	0.600
2	cluster_3	0	0	0.667	0.800
3	cluster_0	1	0	0.333	0.800
4	cluster_2	1	1	0	0.200
5	cluster_2	1	1	0	0.200
6	cluster_3	0	0	0	0
7	cluster_4	0	1	0	0

Cluster IDs indicating categorisation of examples

PerformanceVector

PerformanceVector:
Davies Bouldin: 1.637
Example distribution: 0.212

Parameters

Clustering (k-Means)

☒ add cluster attribute

☐ add as label

☐ remove unlabeled

k ☒ 5

max runs 10

☐ determine good start values ☒

measure types ☒ MixedMeasures

mixed measure MixedEuclideanDista...

max optimization steps 100

☒ use local random seed

local random seed 2020

Experiment with the number of k-means clusters and observe changing performance.

Parameters

Performance DB (Cluster Distance Perfo...

main criterion ☒ Davies Bouldin

☒ main criterion only

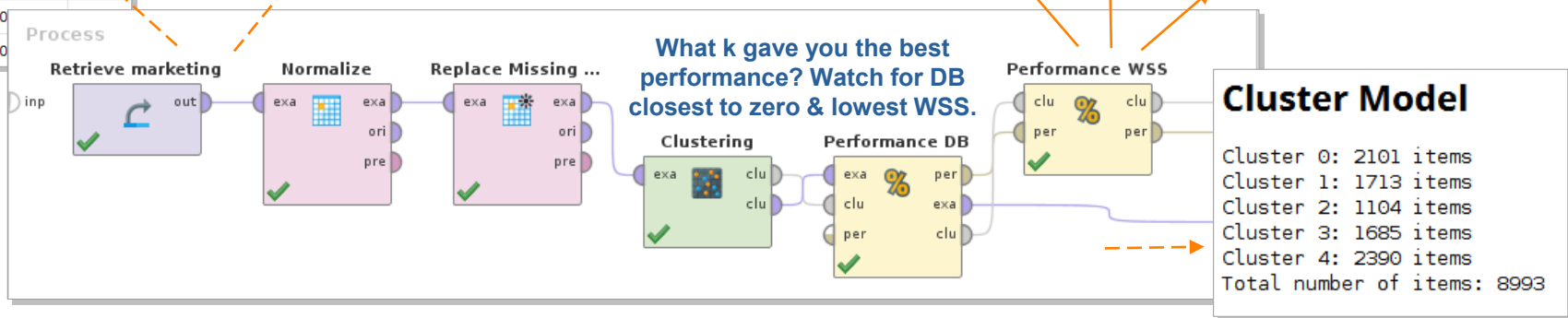
☐ normalize

☒ maximize

Parameters

Performance (Item Distribution Performance)

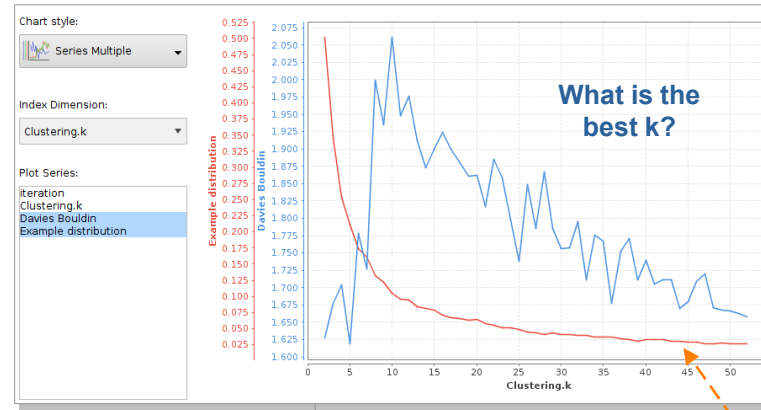
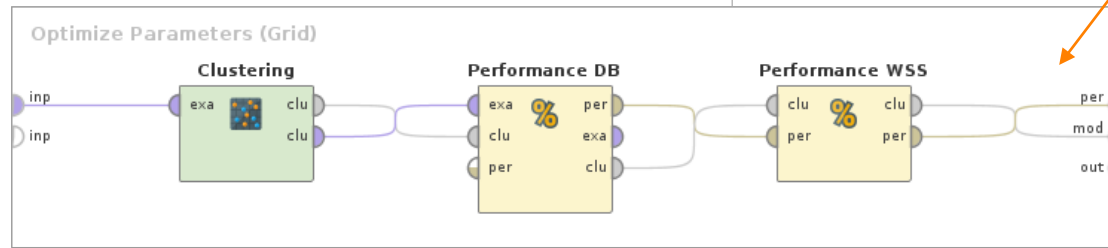
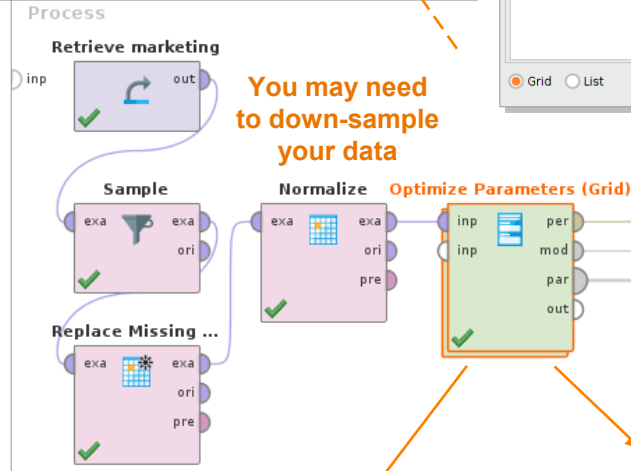
measure SumOfSquares



K-Means Optimisation

Set up an experiment to identify the best number of k-means clusters k within a range 2..52. Save.

Use Cluster Distance Performance (with a single criterion *Davis-Bouldin*) and Item Distribution Performance (with *within SumOfSquares*, or *WSS*). As in the previous process, daisy-chain two performance operators. Ensure that the optimisation grid logs performance of all tracked criteria. Select the best k using Davis-Bouldin index. Also, apply the “elbow” method to WSS.

Parameters

Optimize Parameters (Grid)

Edit Parameter Settings...

error handling: fail on error

☒ log performance

☒ log all criteria

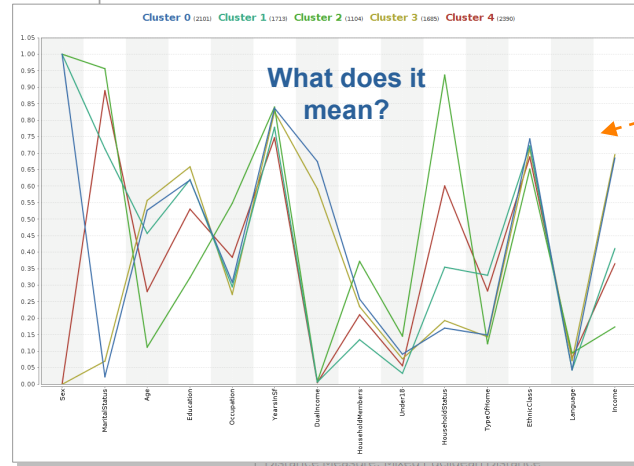
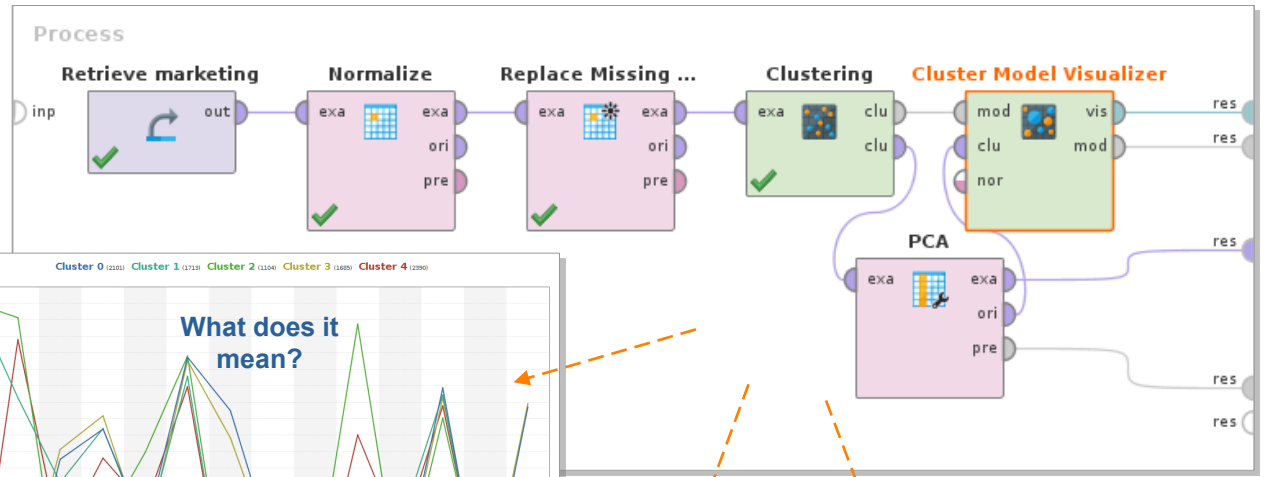
☐ synchronize

☒ enable parallel execution

The grid is likely to get confused when using DB. Why?

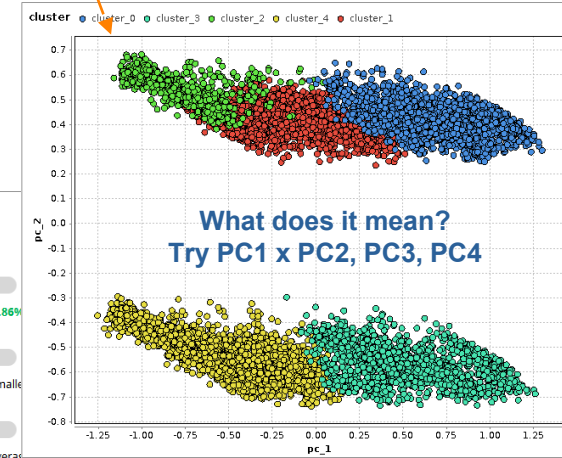
Further Analysis

Take the first clustering process and alter its cluster operator by entering the best k (from optimisation).
Modify the process by adding cluster visualisation.
Run. Interpret results. Save.
Share your insights.



Average Cluster Distance: 0.727
Davies-Bouldin Index: 1.637

Cluster 0	2,101	DualIncome is on average 147.79% larger, MaritalStatus is on average 95.59% smaller, Sex is on average 82.86% smaller
Cluster 1	1,713	DualIncome is on average 97.54% smaller, Sex is on average 82.86% larger, Under18 is on average 56.58% smaller
Cluster 2	1,104	HouseholdStatus is on average 124.05% larger, DualIncome is on average 96.84% smaller, Under18 is on average 56.58% smaller
Cluster 3	1,685	DualIncome is on average 117.56% larger, Sex is on average 100.00% smaller, MaritalStatus is on average 86.44% smaller
Cluster 4	2,390	Sex is on average 100.00% smaller, DualIncome is on average 98.16% smaller, MaritalStatus is on average 75.20% larger



Challenge: Plot and diagnose clusters with PCA. Inspect the cumulative variance plot. Experiment with PCs.