Clustering Selection Management System Report

connect-4.csv – supervised

Thursday 26th November, 2020 – 11:04

Preprocessing

Table 1: Specific Metrics for all Preprocessing steps

Metric	# Rows w/ missings removed	# Converted columns to OHE	# Quantiles for Quant. Scal.	# Non Distinct rows removed
Value	0	42	6755	0

Setup

Hardware

Table 2: Hardware Statistics of the underlying Hardware Setup

Statistic	*Amount of main memory*	*# CPU-Threads*	# CPU-Cores
Value	8.59	4	2

Input Parameters

Table 3: Given *General* Input-Parameter Values

Parameter	Accuracy Efficiency Preference	Prefer Finding arbitrary Cluster Shapes?	Avoid High Effort of (Hyper-) Parameter Tuning?
Value	efficiency	True	False

Table 4: Given *Distance-Metric-based* Input-Parameter Values

Parameter	Find Compact or Isolated Clusters?	Ignore Magnitude and Rotation?	Measure Distribution Differences?	Grid-based Distance?
Value	True	False	False	False

Metadata

Table 5: *General* Profiled Metadata Results regarding the Dataset

Statistic	*#Rows*	*#Columns*	#Classes	# Missing Values
Value	67557	123	3	0

Table 6: *Further* Profiled Metadata Results regarding the Dataset

Statistic	*Outlier %*	*High Correlation %*	Class Std. Deviation
Value	0.0038	0.0063	19683.044378347575

Selection Steps

Table 7: Listing of all CSMS Iterations

Iteration	Selected Algorithm	Selection-Score	Tuned (Hyper-) Parameters	Accuracy of Sampling
Iteration 1	nearest_centroid	8.59	distance = manhattan	0.45
Iteration 2	svc_sgd	8.17		0.66
Iteration 3	radius_neighbors	6.32	distance = manhattan, radius = 1000	0.63
Iteration 4	svc	5.77	degree = 9	0.65
Iteration 5	knn	4.74	distance = manhattan, k = 4	0.58
Iteration 6	nca	2.53	distance = manhattan, k = 8	0.66

Results

Table 8: Final Clustering Result

Algorithm	Tuned (Hyper-) Parameters	Reached Accuracy	Total CPU-Runtime of the CSMS
svc_sgd		0.75	23.54s