RapidMiner Homework

Mehrdad Mohammadian   39700248

Problem

Detect Phishing Websites - Classification & Clustering

Data

Available in Kaggle: https://www.kaggle.com/eswarchandt/phishing-website-detector

Paper: http://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf

-	Value	Info
Number of samples	11053	Untitled
Number of Features	30	Untitled
Type of Fetures	Integer	Untitled
Labels	phishing (1) / non-phishing (-1)	Untitled

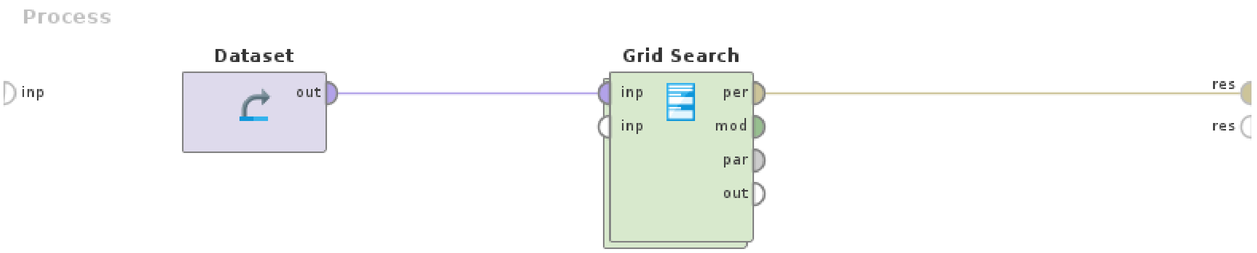
RapidMiner Process

1- Classification Process

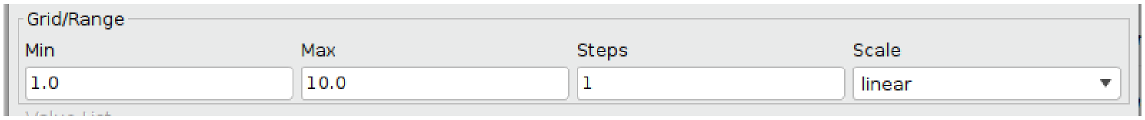
3 layer

Layer 1: Dataset & Grid Search

Dataset: Set Role For Label, In Import Dataset Step

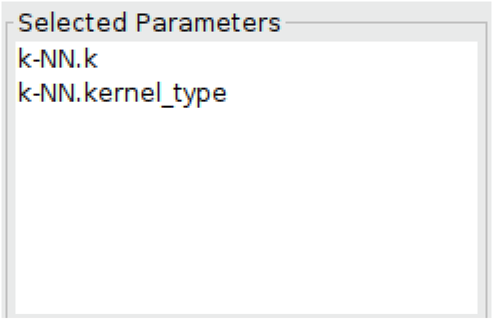


k:



Grid/Range

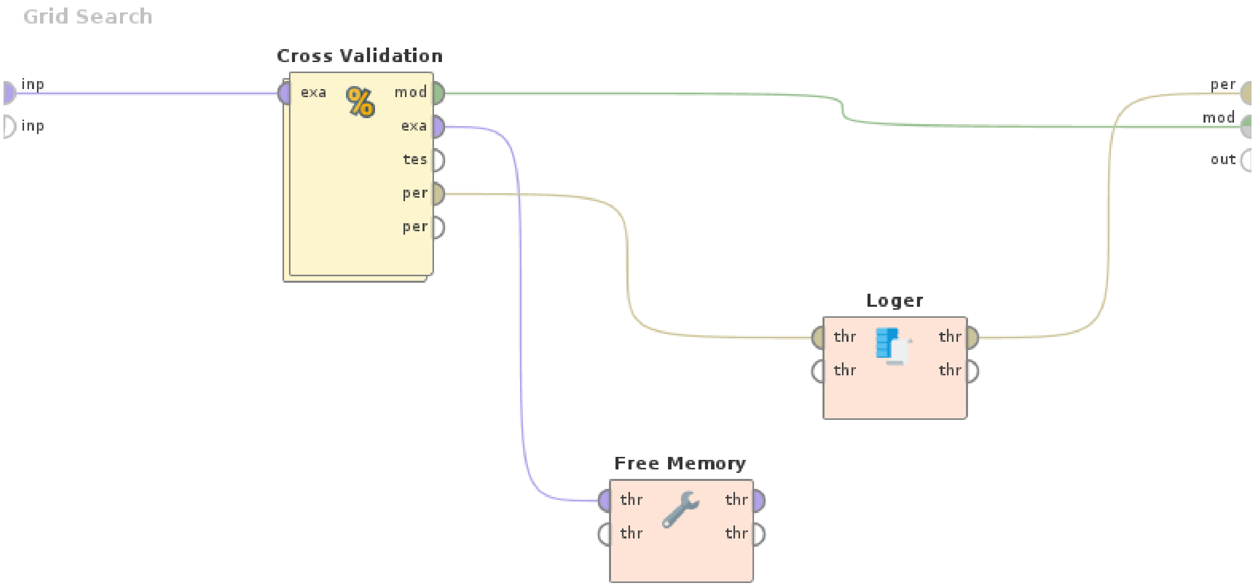
Min	Max	Steps	Scale
1.0	10.0	1	linear



Selected Parameters

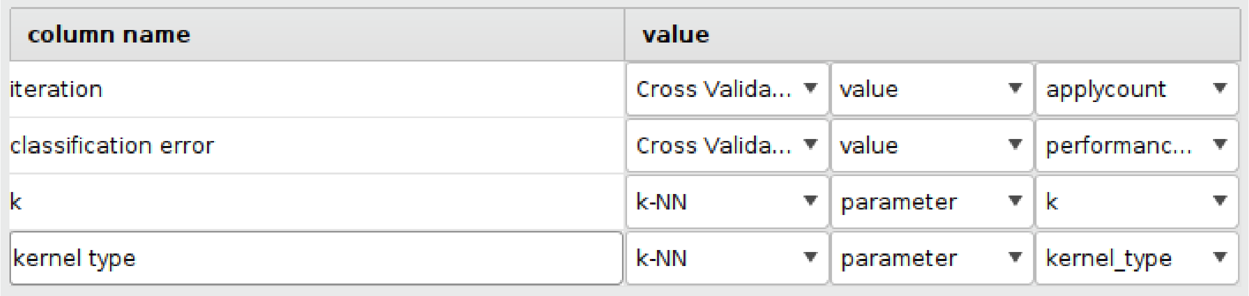
k-NN.k
k-NN.kernel_type

Layer 2: Cross Validation & Loger & Free Memory



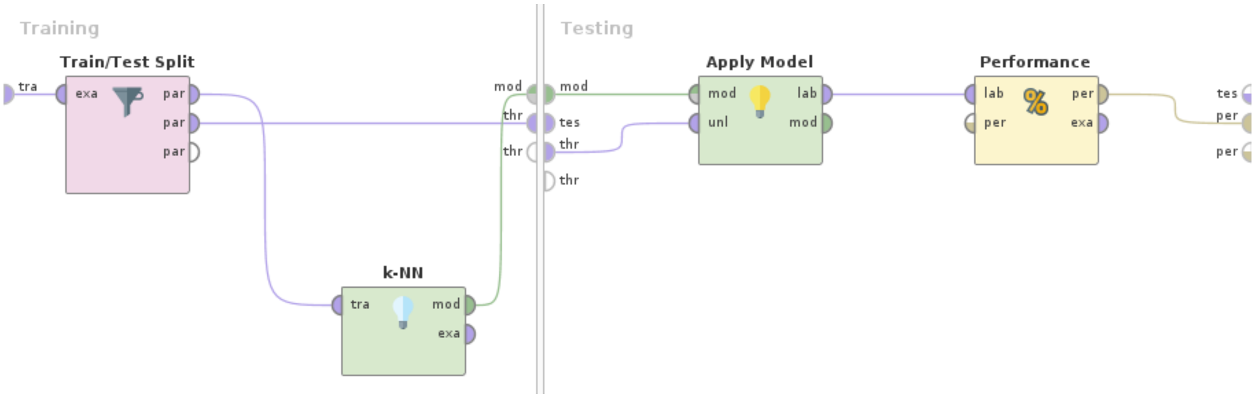
Cross Validation: 10 Folds

Loger:



column name	value		
iteration	Cross Valida... ▼	value ▼	applycount ▼
classification error	Cross Valida... ▼	value ▼	performanc... ▼
k	k-NN ▼	parameter ▼	k ▼
kernel type	k-NN ▼	parameter ▼	kernel_type ▼

Layer 3: Split Data & KNN Model & Performance



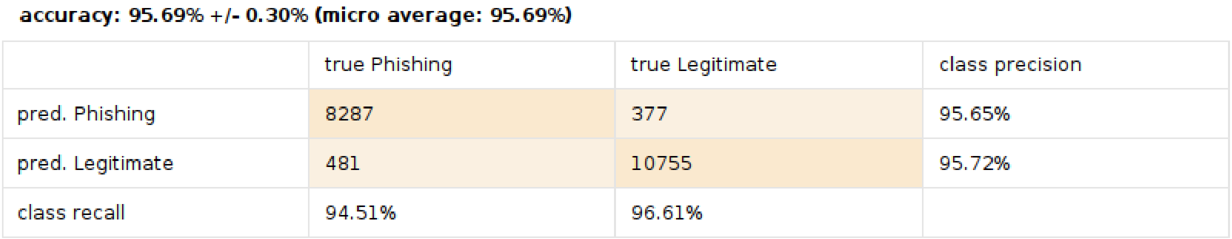
Split Data :   Test 0.2, Train 0.8

Results:

Accuracy: 95.69 %

Classification Error: 4.31% +/- 0.30%

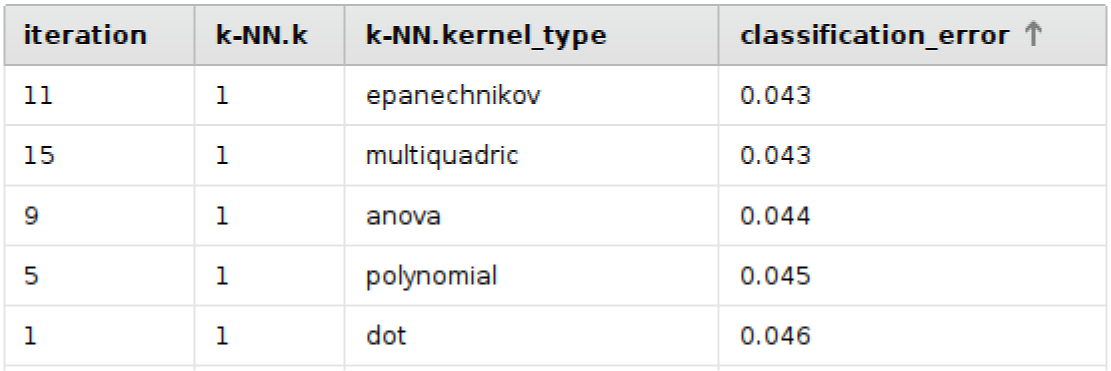
Confusion Matrix:



accuracy: 95.69% +/- 0.30% (micro average: 95.69%)

	true Phishing	true Legitimate	class precision
pred. Phishing	8287	377	95.65%
pred. Legitimate	481	10755	95.72%
class recall	94.51%	96.61%	

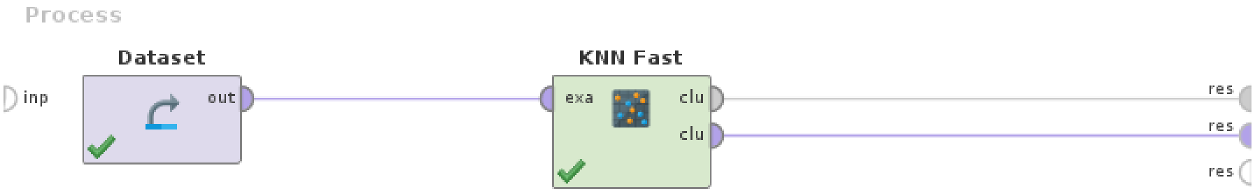
Top 5 Models:

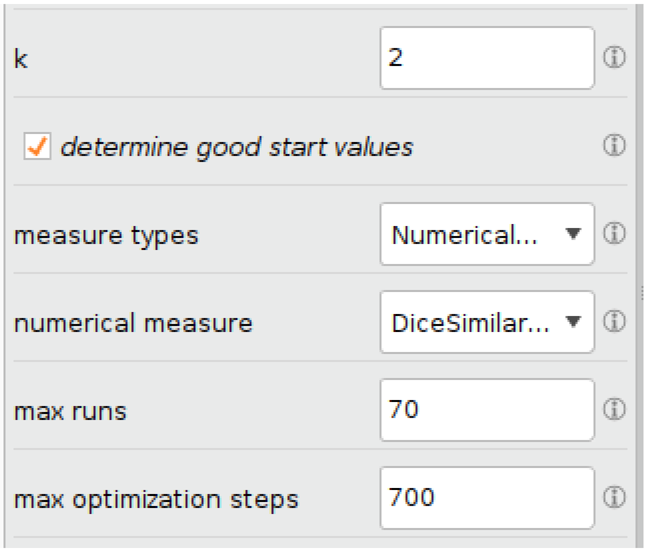


iteration	k-NN.k	k-NN.kernel_type	classification_error ↑
11	1	epanechnikov	0.043
15	1	multiquadric	0.043
9	1	anova	0.044
5	1	polynomial	0.045
1	1	dot	0.046

2- Clustering Process

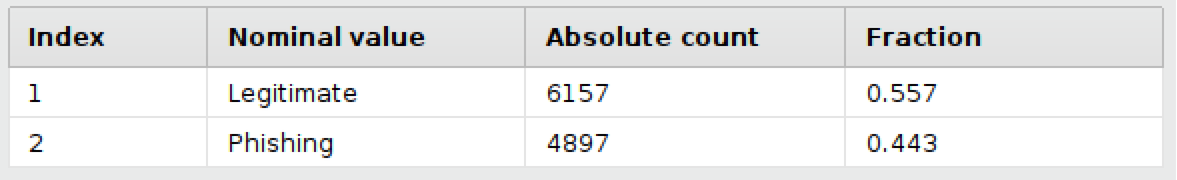
1 layer



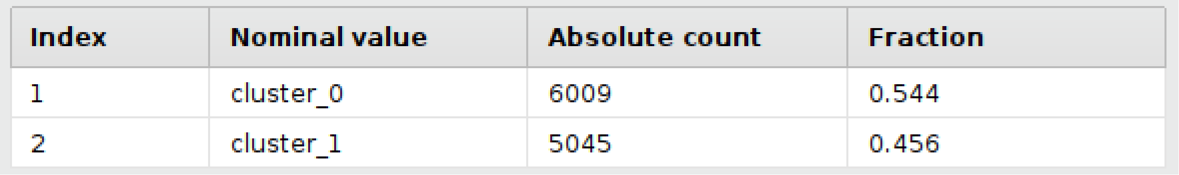


k	2
☑ *determine good start values*	
measure types	Numerical...
numerical measure	DiceSimilar...
max runs	70
max optimization steps	700

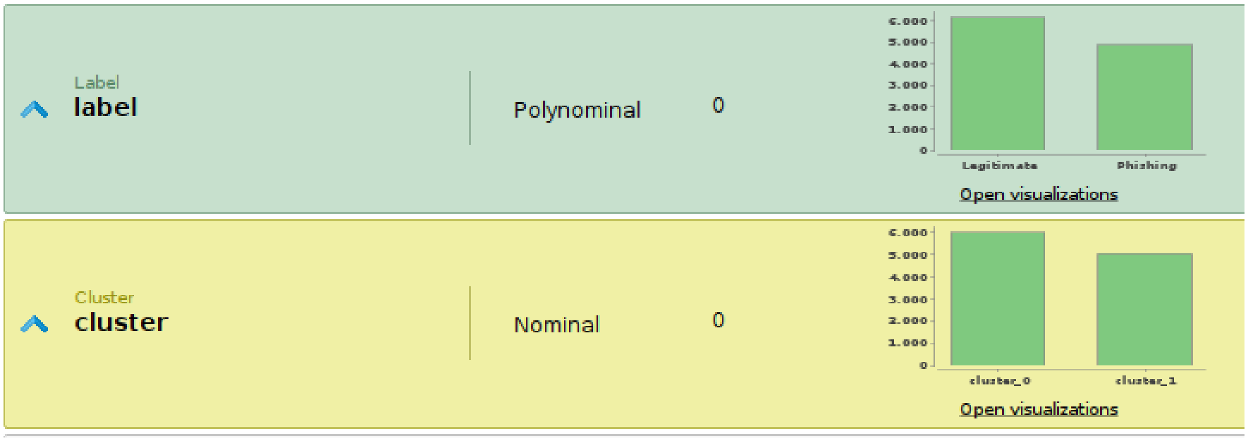
Compare Results:



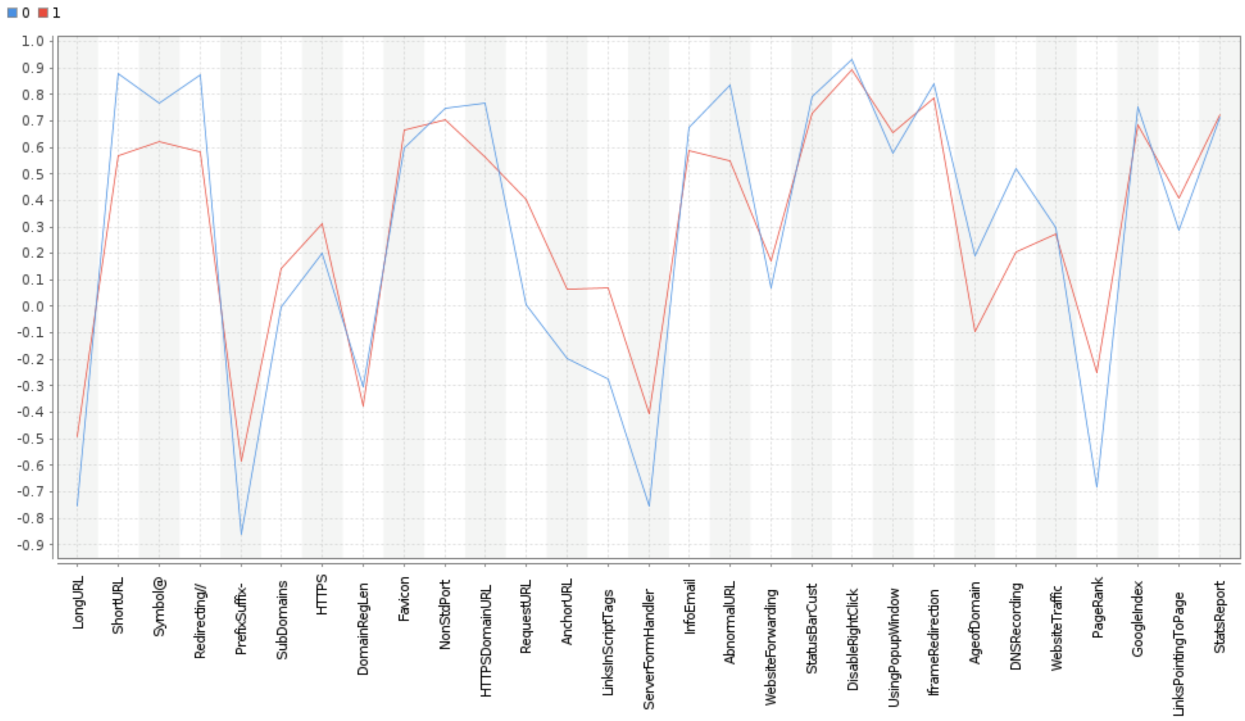
Index	Nominal value	Absolute count	Fraction
1	Legitimate	6157	0.557
2	Phishing	4897	0.443



Index	Nominal value	Absolute count	Fraction
1	cluster_0	6009	0.544
2	cluster_1	5045	0.456



Label label	Polynominal	0	Open visualizations
Cluster cluster	Nominal	0	Open visualizations


🟦 0 🟥 1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
LongURL
ShortURL
Symbol@
Redirecting//
PrefixSuffix-
SubDomains
HTTPS
DomainRegLen
Favicon
NonStdPort
HTTPSDomainURL
RequestURL
AnchorURL
LinksInScriptTags
ServerFormHandler
InfoEmail
AbnormalURL
WebsiteForwarding
StatusBarCust
DisableRightClick
UsingPopupWindow
IframeRedirection
AgeofDomain
DNSRecording
WebsiteTraffic
PageRank
GoogleIndex
LinksPointingToPage
StatsReport

The End