

Problem Description

- A complex modern **semi-conductor manufacturing process** is normally under consistent surveillance via the monitoring of signals/variables collected from sensors and or process measurement points.
- However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain **a combination of useful information, irrelevant information as well as noise.**

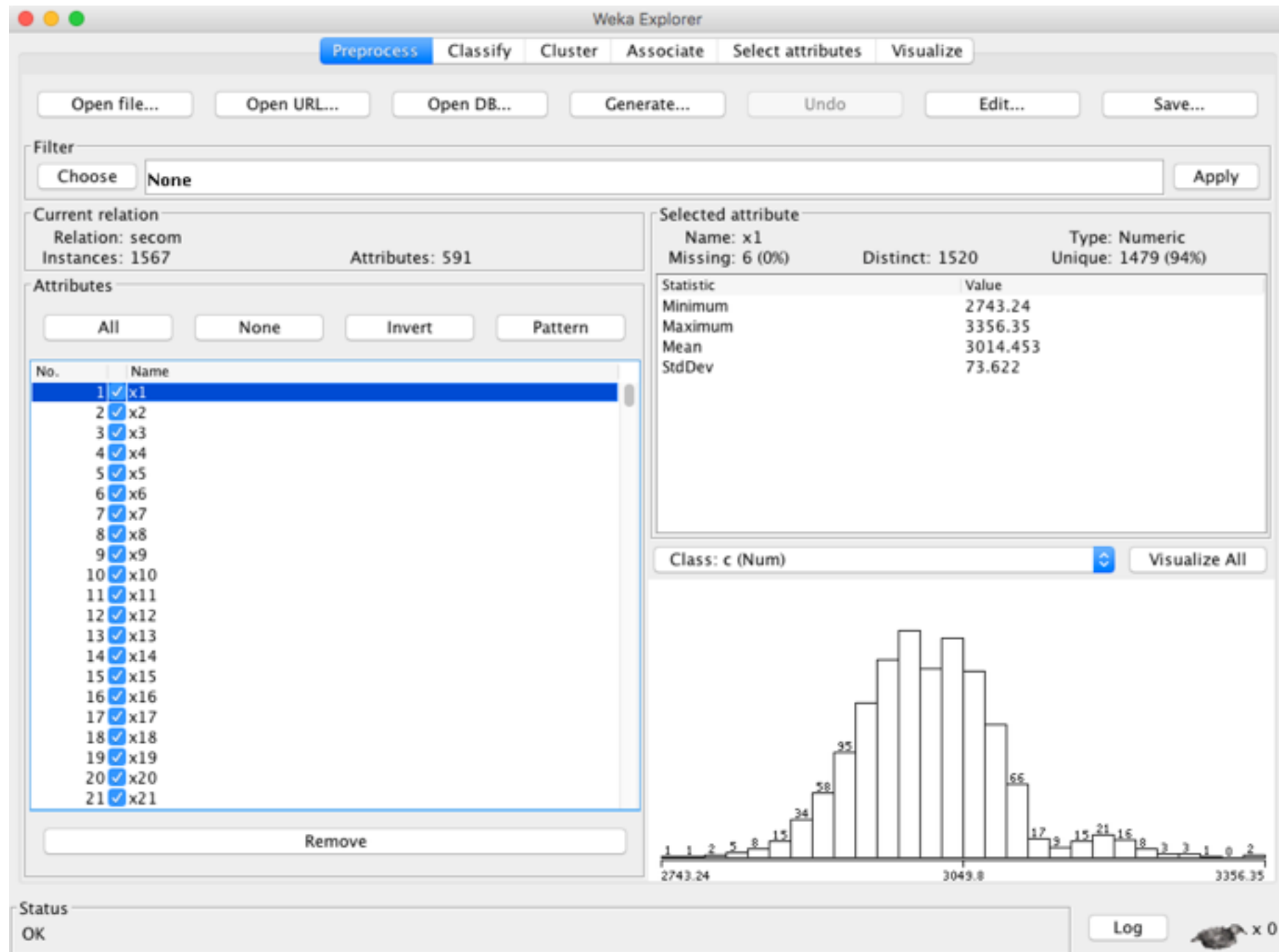
Problem Description

- Engineers typically have a much larger number of signals than are actually required. **If we consider each type of signal as a feature**, then feature selection may be applied to identify the **most relevant signals**.
- The Process Engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learning and reduce the per unit production costs.

SECOM Dataset

- SECOM Dataset: 1567 examples 591 features, 104 fails
- There are missing values;
- Where -1 corresponds to a pass and 1 corresponds to a fail and the data time stamp is for that specific test point.
- <https://archive.ics.uci.edu/ml/datasets/SECOM>

Features (variables)



ARFF File

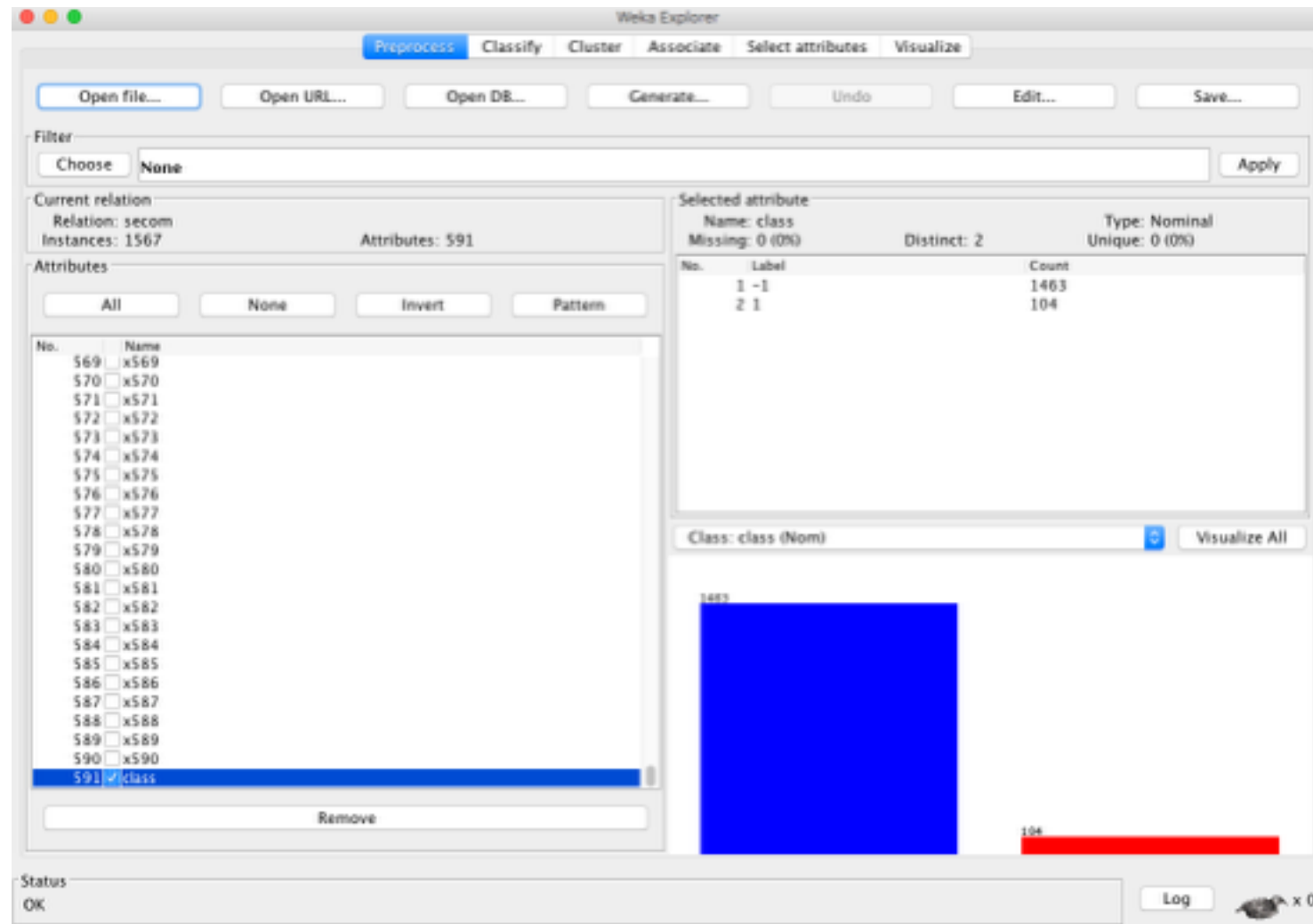
The variable **class** is the binary.

@attribute class $\{-1, 1\}$

-1 = Pass
1 = Fail

```
@attribute x588 numeric
@attribute x589 numeric
@attribute x590 numeric
@attribute class {-1,1}

@data
3030.93,2564,2187.7333,1411.1265,1.3602,100,97.61
455,202.4396,0,7.9558,414871,10.0433,968,192.3963
751,0.0055,1770,2040,61,0.0000,0.0000,0.1000,0.511
```



Algorithms used

- Naive Bayes
- TAN
- IB1
- Idk
- RIPPER
- ID3
- C4.5 (J48)
- Logistic

Measures

=== Stratified cross-validation ===
 === Summary ===

Correctly Classified Instances	1012	64.582 %
Incorrectly Classified Instances	555	35.418 %
Kappa statistic	0.0077	
Mean absolute error	0.3534	
Root mean squared error	0.59	
Relative absolute error	283.9367 %	
Root relative squared error	237.0124 %	
Total Number of Instances	1567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.666	0.644	0.936	0.666	0.778	0.505	-1
	0.356	0.334	0.07	0.356	0.118	0.497	1
Weighted Avg.	0.646	0.624	0.878	0.646	0.735	0.505	

=== Confusion Matrix ===

a	b	<-- classified as
975	488	a = -1
67	37	b = 1

		prediction outcome		total
		p	n	
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	

It is possible to see that 975 are True Positive, 488 False Negative, 67 False Positive and 37 True Negative.

Naive Bayes with all variables.

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      1012      64.582 %
Incorrectly Classified Instances    555      35.418 %
Kappa statistic                    0.0077
Mean absolute error                 0.3534
Root mean squared error             0.59
Relative absolute error             283.9367 %
Root relative squared error         237.0124 %
Total Number of Instances          1567

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.666	0.644	0.936	0.666	0.778	0.505	-1
	0.356	0.334	0.07	0.356	0.118	0.497	1
Weighted Avg.	0.646	0.624	0.878	0.646	0.735	0.505	

```

=== Confusion Matrix ===

```

a	b	<-- classified as
975	488	a = -1
67	37	b = 1

		prediction outcome		
		p	n	total
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	

It is possible to see that 975 are True Positive, 488 False Negative, 67 False Positive and 37 True Negative. 64.58% Correctly Classified Instances.

TAN with all variables.

Where is it?

% Correctly Classified Instances.

IB1 with all variables.

Classifier

Choose IB1

Test options

☐ Use training set

☐ Supplied test set Set...

☒ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 19:14:10 - bayes.NaiveBayes
- 19:26:27 - trees.RandomForest
- 19:28:31 - lazy.IB1

Classifier output

151	2:1	2:1	0	*1
152	2:1	2:1	0	*1
153	2:1	1:-1	+ *1	0
154	2:1	1:-1	+ *1	0
155	2:1	1:-1	+ *1	0
156	2:1	2:1	0	*1

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	1297	82.7696 %
Incorrectly Classified Instances	270	17.2304 %
Kappa statistic	0.8841	
K&B Relative Info Score	-148577.2796 %	
K&B Information Score	-528.2089 bits	-0.3371 bits/instance
Class complexity order 0	551.9814 bits	0.3523 bits/instance
Class complexity scheme	289980 bits	185.0542 bits/instance
Complexity improvement (Sf)	-289428.0186 bits	-184.702 bits/instance
Mean absolute error	0.1723	
Root mean squared error	0.4151	
Relative absolute error	138.4386 %	
Root relative squared error	166.7511 %	
Total Number of Instances	1567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.868	0.74	0.943	0.868	0.904	0.564	-1
	0.26	0.132	0.123	0.26	0.167	0.564	1
Weighted Avg.	0.828	0.7	0.888	0.828	0.855	0.564	

=== Confusion Matrix ===

a	b	<-- classified as
1270	193	a = -1
77	27	b = 1

82.76% Correctly Classified Instances.

Idk with all variables.

Where is it?

% Correctly Classified Instances.

RIPPER with all variables.

Where is it?

% Correctly Classified Instances.

ID3 with all variables.

Where is it?

% Correctly Classified Instances.

C4.5 (J48) with all variables.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'J48 -C 0.25 -M 2'. The test options are set to 'Cross-validation' with 10 folds. The classifier output shows a stratified cross-validation summary with 1417 correctly classified instances (90.4276%) and 150 incorrectly classified instances (9.5724%). The result list on the left shows 'trees.J48' as the selected model. The status bar at the bottom indicates 'OK'.

Classifier output

Instance	Actual	Predicted	Confidence	Weight
151	2:1	2:1	0	+1
152	2:1	2:1	0	+1
153	2:1	1:-1	+	*1
154	2:1	1:-1	+	*0.998
155	2:1	1:-1	+	*0.998
156	2:1	1:-1	+	*1

=== Stratified cross-validation ===
=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	1417	90.4276 %
Incorrectly Classified Instances	150	9.5724 %
Kappa statistic	0.0796	
K&B Relative Info Score	-47936.3228	%
K&B Information Score	-170.419	bits
Class complexity order 0	551.9814	bits
Class complexity scheme	63902.9649	bits
Complexity improvement (Sf)	-63350.9835	bits
Mean absolute error	0.1112	
Root mean squared error	0.3016	
Relative absolute error	89.3181	%
Root relative squared error	121.1626	%
Total Number of Instances	1567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.961	0.894	0.938	0.961	0.949	0.57	-1
	0.106	0.039	0.162	0.106	0.128	0.57	1
Weighted Avg.	0.904	0.837	0.886	0.904	0.895	0.57	

=== Confusion Matrix ===

a	b	← classified as
1406	57	a = -1
93	11	b = 1

90.42% Correctly Classified Instances.

Logistic with all variables.

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier chosen is 'Logistic -R 1.0E-8 -M -1'. The test options are set to 'Cross-validation' with 10 folds. The result list on the left shows '19:38:19 - functions.Logistic' selected. The classifier output on the right displays the following data:

Classifier output

Instance	Class	Actual	Predicted	Confidence
151	2:1	2:1	0	*1
152	2:1	2:1	0	*1
153	2:1	2:1	0	*1
154	2:1	1:-1	+	*1
155	2:1	1:-1	+	*1
156	2:1	2:1	0	*1

=== Stratified cross-validation ===
=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	1268	80.919 %
Incorrectly Classified Instances	299	19.081 %
Kappa statistic	0.0394	
K&B Relative Info Score	-185402.9384	%
K&B Information Score	-659.1283	bits
Class complexity order 0	551.9814	bits
Class complexity scheme	23839.1024	bits
Complexity improvement (Sf)	-23287.121	bits
Mean absolute error	0.1905	
Root mean squared error	0.4327	
Relative absolute error	153.0268	%
Root relative squared error	173.8323	%
Total Number of Instances	1567	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.852	0.788	0.938	0.852	0.893	0.557	-1
	0.212	0.148	0.092	0.212	0.128	0.564	1
Weighted Avg.	0.809	0.746	0.882	0.809	0.842	0.557	

=== Confusion Matrix ===

a \ b	-1	1
-1	1246	217
1	82	22

Logistic classifier results summary:

- Correctly Classified Instances: 1268 (80.919%)
- Incorrectly Classified Instances: 299 (19.081%)
- Kappa statistic: 0.0394
- Mean absolute error: 0.1905
- Root mean squared error: 0.4327
- Relative absolute error: 153.0268%
- Root relative squared error: 173.8323%
- Total Number of Instances: 1567

Detailed Accuracy By Class:

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
-1	0.852	0.788	0.938	0.852	0.893	0.557
1	0.212	0.148	0.092	0.212	0.128	0.564

Confusion Matrix:

a \ b	-1	1
-1	1246	217
1	82	22

80.91% Correctly Classified Instances.

Performance

	All variables	FSS1	FSS2	Wrapper
Naive Bayes	64.58%			
TAN				
IB1	82.76%			
IBK				
RIPPER				
ID3				
C4.5 (J48)	90.42%			
Logistic	80.91%			

Conclusions

- We can see....

Thanks

- Caio Fernandes Moreno
caio.fmoreno@alumnos.upm.es