Predictive Quality Data Analyses

Reference Results for PredQuality_Data1.xlsx

ICNAP Hackathon, 25-27 October 2019

IconPro GmbH

Overall Conclusion

- Analysis A
  - We assume that relationships between Assembly Data and Initial Inspection data exist in high dimensional space
  - We could derive the 5 most important influencing variables
  - The derived relationships are fuzzy and overfit in favor of the non-outlier class
- Analysis B:
  - We could show that relationships between Initial Inspection Data and Final Inspection data exist in high dimensional space
  - We could derive the 5 most important influencing variables
  - The derived relationships are much clearer and do not overfit in favor of any class
- Analysis C:
  - Relations between Assembly Data and Final Inspection Data data seem to exist in high dimensional space
  - We did not find significant relationships that can reasonably derived from the data

- In summary:
  - Significant relationships in between Assembly Data and Initial Inspection Data as well as in between Initial and Final Inspection Data are to be expected
  - Clear relationships could be derived in between Initial Inspection Data and Final Inspection Data when it comes to outlier classification
  - The data-modelling was partly improved by more data / oversampling, i.e. especially more data for “outliers”
    (s. SMOTE: Synthetic Minority Over-sampling Technique)
  - So far, no regression model could be derived
  - Outlier classification may help to predict bad-quality-parts and reduce inspection efforts by just inspecting aforementioned parts!

ICON PRO

Data Pre-processing

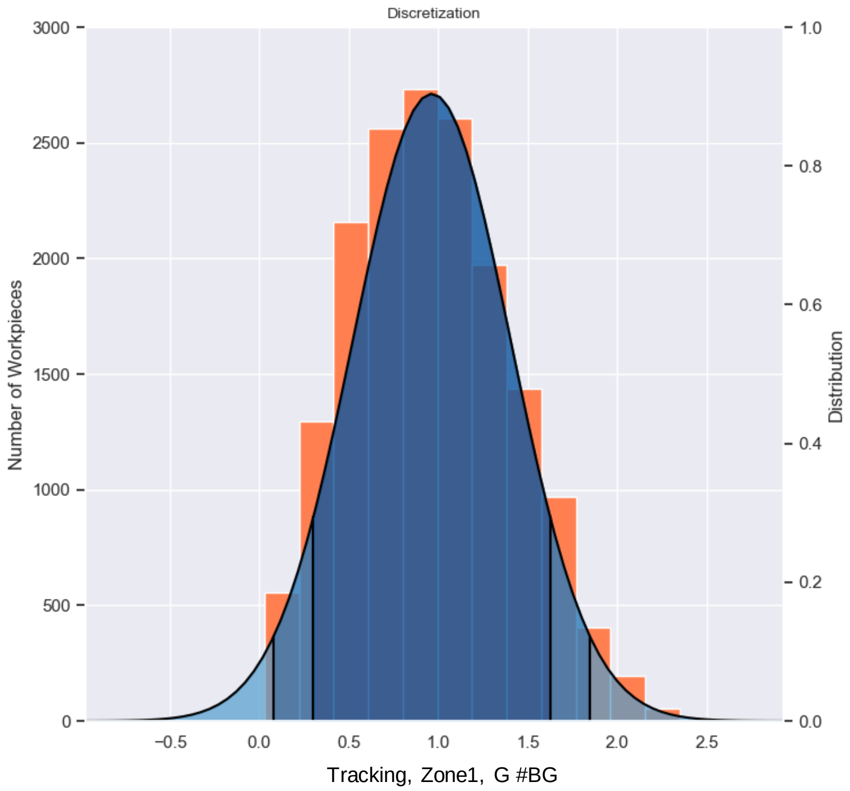
- Pre-processing preserves occurrence of all characteristics from the excel sheets
- For analysis B (initial to final inspection) and C (assembly to final inspection) data for some workpieces were excluded, as their initial inspection data is equal to the final inspection data, or there was no final inspection data

- Analysis inputs:
  - Normalized the angle characteristics to the range of [0, 360] by modulo operation
  - For neural network approaches, non-angle characteristics were also normalized by simple z-score manipulation
- Analysis outputs:
  - In early analysis iterations the original numerical values of each output characteristic were used
  - This approach was not adequate as the numerical target values mostly concentrate around a specific value (therefore skipped regression analyses)
  - As a mean of complexity reduction, the analysis outputs were discretized, such that numerical ranges for each output characteristic were grouped by z-scores (i.e., defines by how many standard deviations a value is away from the sample mean, see next slide for example)

Output Data Pre-processing – Discretization Details

- For complexity reduction from numerical range to discrete space (Assumption of underlying normal distribution)
- Such discretization allows for outlier classification analysis
- Class definitions, for example (see coloured area under curve):
  - Class 0: Distance of up to ± 1.5 STDs away from sample mean
  - Class 1: Distance above ± 1.5 and up to ± 2 STDs away from sample mean
  - Class 2: Distance above ± 2 STDs away from sample mean
  - When applied, discretization will be given in form of:

    [0: <= ± 1.5 | 1: <= ± 2.0 | 2: > 2.0]
- Oversampling is partly used for data analyses to compensate for unbalanced class sizes in given data



Data Analysis B – Exemplary Results (Confusion Matrices)

Predict / Actual	0	1	Sensitivity / Specificity
0	TP	FP	TP/(TP+FP)% **[SN]**
1	FN	TN	TN/(TN+FN)% **[SP]**
Total	*Output Char.*		X%

Legend

- As Final NG class is underrepresented, a second approach aims to predict either TOL A or not TOL A

	Tol A	Tol B	Final NG	Accuracies
Tol A	2250	616	15	78.10%
Tol B	96	260	3	72.42%
Final NG	2	2	3	42.86%
Total	Final result #AV			64.45%

	Tol A	not Tol A	Accuracies
Tol A	2440	422	84.66%
not Tol A	111	254	69.59%
Total	Final result #AV		77.13%

Discretization:
[0: <= ± 2.5 | 1: <= ± 2.5]

ICON PRO

Data Analysis B – Exemplary Results (Confusion Matrices)

Predict / Actual	0	1	Sensitivity / Specificity
0	TP	FP	TP/(TP+FP)% **[SN]**
1	FN	TN	TN/(TN+FN)% **[SP]**
Total	*Output Char.*		X%

Legend

	0	1	
0	2467	631	79.63%
1	37	112	75.17%
Total	weight #BO		77.40%

	0	1	
0	2203	928	70.36%
1	29	87	75.00%
Total	weight #BQ		72.68%

	0	1	
0	2530	607	80.65%
1	22	88	80.00%
Total	Tracking, Zone1 #BS		80.33%

	0	1	
0	2458	679	78.36%
1	26	84	76.36%
Total	Tracking, Zone2 #BT		77.36%

	0	1	
0	2181	915	70.45%
1	63	88	58.28%
Total	Tracking, Zone3 #BU		64.37%

	0	1	
0	2069	1029	66.79%
1	34	115	77.18%
Total	Tracking, Zone4 #BV		72.00%

Discretization:
[0: <= ± 2.5 | 1: <= ± 2.5]

Data Analysis A – Exemplary Results (Confusion Matrices)

Predict / Actual	0	1	2	Accuracies
0	A	B	C	A/(A+B+C)%
1	C	E	F	E/(C+E+F)%
2	G	H	J	J/(G+H+J)%
Total	*Output Char.*			Mean%

Legend

- Results for all output characteristics were of similar quality, thus here exemplarily just one (Tracking, Zone1, G #BG) is shown:

All input characteristics

	0	1	2	Accuracies
0	3227	51	23	97.75%
1	77	7	2	8.14%
2	5	0	1	16.67%
Total	Tracking, Zone1, G #BG			40.85%

Only best five input characteristics

	0	1	2	Accuracies
0	2571	489	235	78.02%
1	47	33	13	35.48%
2	0	3	2	20.00%
Total	Tracking, Zone1, G #BG			44.33%

All but best five input characteristics

	0	1	2	Accuracies
0	3264	21	12	98.90%
1	91	2	0	2.15%
2	2	0	1	33.33%
Total	Tracking, Zone1, G #BG			44.68%

Discretization:
[0: <= ± 2.0 | 1: <= ± 3.0 | 2: > 3.0]

ICON PRO

Data Analysis A – Exemplary Results (Confusion Matrices)

	0	1	2	Acc.
0	2609	524	108	80.50%
1	82	31	4	26.50%
2	18	11	6	17.14%
Total	Tracking, Zone4, G #BG			41.38%

	0	1	2	Acc.
0	2848	287	119	87.52%
1	76	18	6	18.00%
2	29	4	6	15.38%
Total	Tracking, Zone4, G #BG			40.30%

	0	1	2	Acc.
0	2953	233	67	90.78%
1	90	16	4	14.55%
2	24	1	5	16.67%
Total	Tracking, Zone4, G #BG			40.67%

	V	Asc.	Desc.	Flat	Acc.
V	420	479	288	158	31.23%
Asc.	334	446	215	172	38.22%
Desc.	224	254	197	144	24.05%
Flat	7	31	15	9	14.52%
Total	Shape #BK				27.01%

	V	Asc.	Desc.	Flat	Acc.
V	575	414	256	114	42.31%
Asc.	412	441	272	115	35.56%
Desc.	267	192	208	80	27.84%
Flat	10	20	9	8	17.02%
Total	Shape #BK				30.68%

	V	Asc.	Desc.	Flat	Acc.
V	517	379	401	26	39.08%
Asc.	433	388	346	36	32.25%
Desc.	313	202	259	23	32.25%
Flat	17	31	22	0	0.00%
Total	Shape #BK				25.90%

All input characteristics	**Only best five input characteristics**	All but best five input characteristics

Discretization:
[0: <= ± 2.0 | 1: <= ± 3.0 | 2: > 3.0]

ICON PRO

Data Analysis C – Exemplary Results (Confusion Matrices)

- The best results are achieved on HSB final result classification accuracy based on assembly characteristics as seen below

	Tol A	not Tol A	Accuracies
Tol A	1982	904	68.68%
not Tol A	199	162	44.88%
Total	Final result #AV		56.78%

- Results are not significantly good, as class Tol A cannot be recognized well enough (overfitting)

These were exemplary reference results from us for PredQuality_Data1.xlsx only.
Can you achieve the same or maybe you can do better, i.e. predict outputs more accurate?

Maybe there are even better relationships hidden in PredQuality_Data2.xlsx?
Maybe it makes sense to merge the data sets?

Maybe you find continuous relationships for continuous outputs or
a better discretization of the continuous output than we did?

It is not an easy task, therefore we provided these exemplary results as a reference.
As this is real industrial field data, there will be no perfect solution

Have fun!

THANKS!

Any questions? You can find me at

Markus.Ohlenforst@iconpro.com