Normalizing data for outlier detection

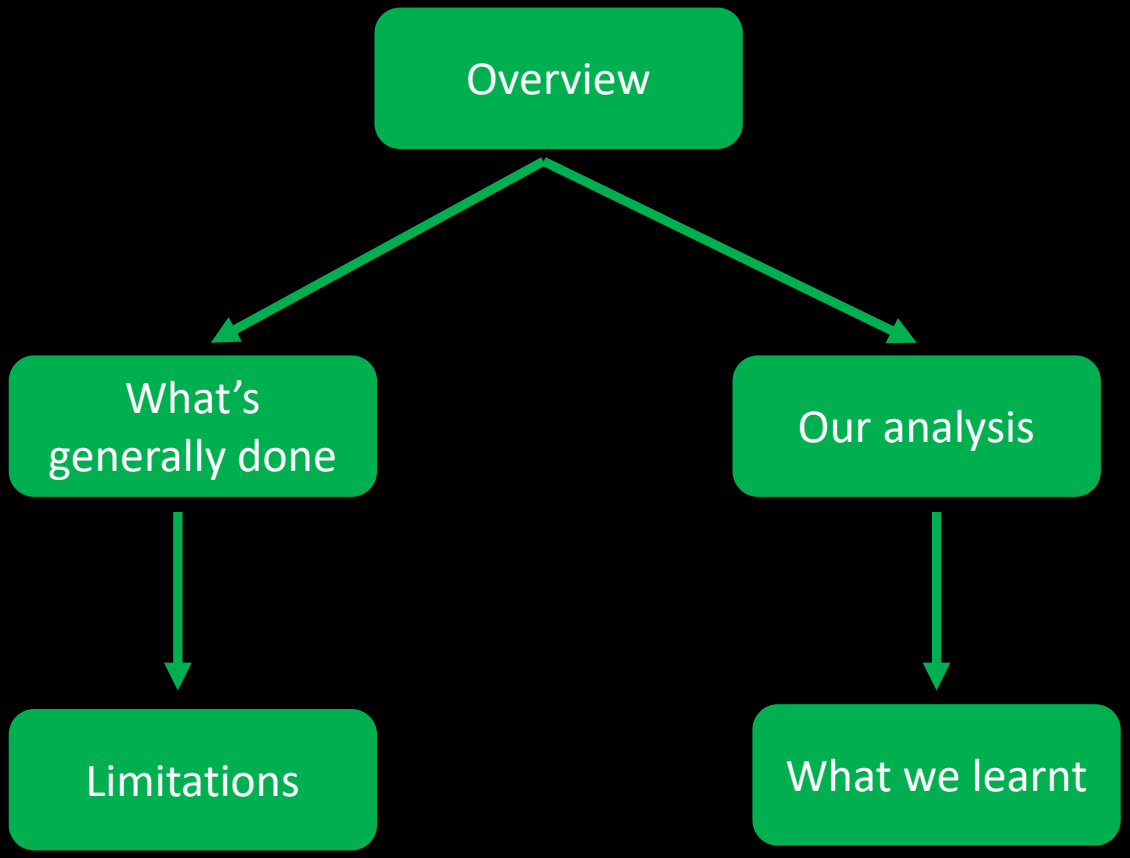
Sevvandi Kandanaarachchi

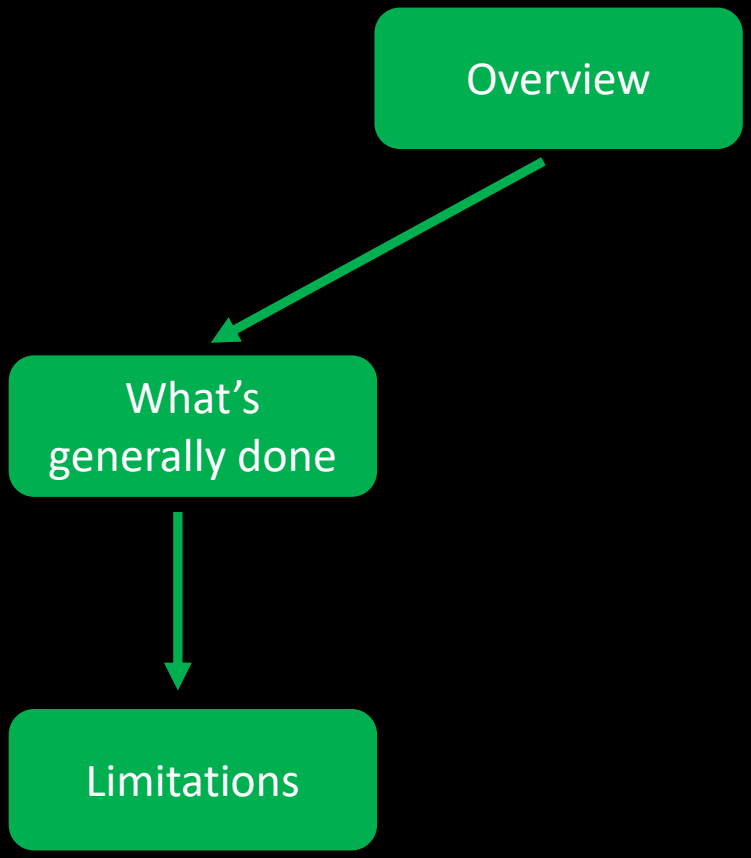
useR2018

Wednesday 11 July 2018

Joint work with

- Rob Hyndman
- Kate Smith-Miles
- Añdres Muñoz-Acosta


Overview
What's generally done
Our analysis
Limitations
What we learnt


Overview
What's generally done
Limitations

Outlier detection

- Min-max normalization
- All columns are normalized to $[0,1]$

Other normalization/standardization

- Mean-SD
- Median-IQR
- Median-MAD (median absolute deviation)
- Apart from Min-Max

OK. So what’s the problem?

Well . . .
Performance may suffer . . .

LOF - An example

LOF – local outlier factor
(Breunig *et. al*.)

Filename	LOF (mean-sd)	LOF (median-iqr)	LOF (median-mad)	LOF (min-max)
Dataset 1 (abalone)	0.65	0.65	0.64	0.66
Dataset 2 (annthyroid)	0.70	0.86	0.90	0.56

Performance – Area under ROC curve

But WHY does it affect performance?

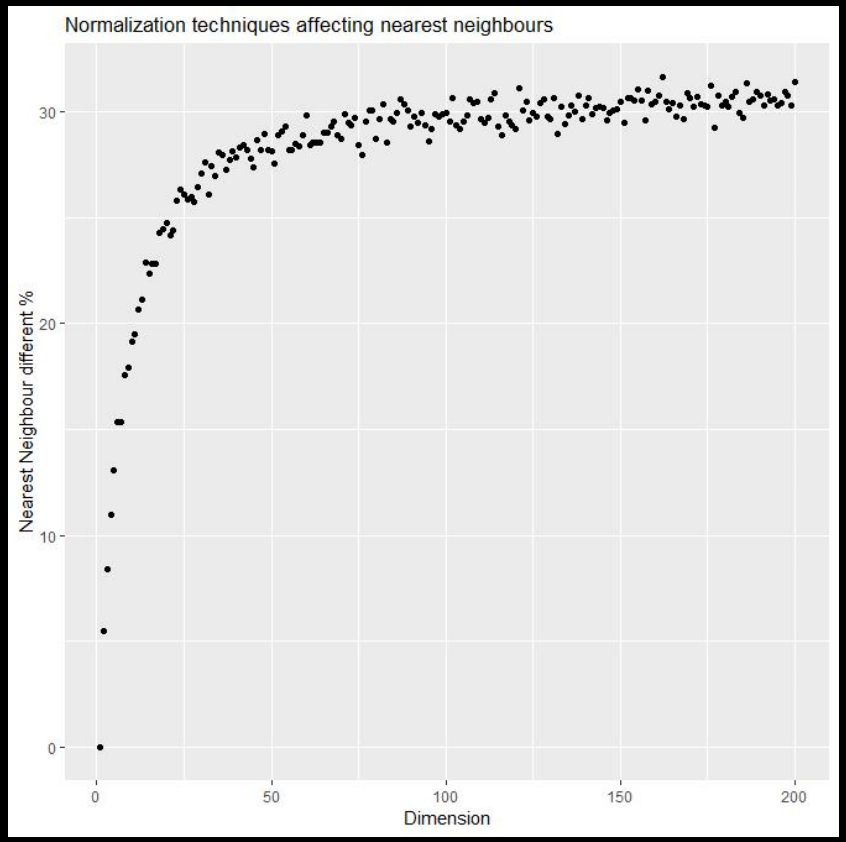
- Each axis is squashed/stretched differently
- Normalization changes distances
- Nearest neighbour structure changes
- Changes densities

Change of nearest neighbours

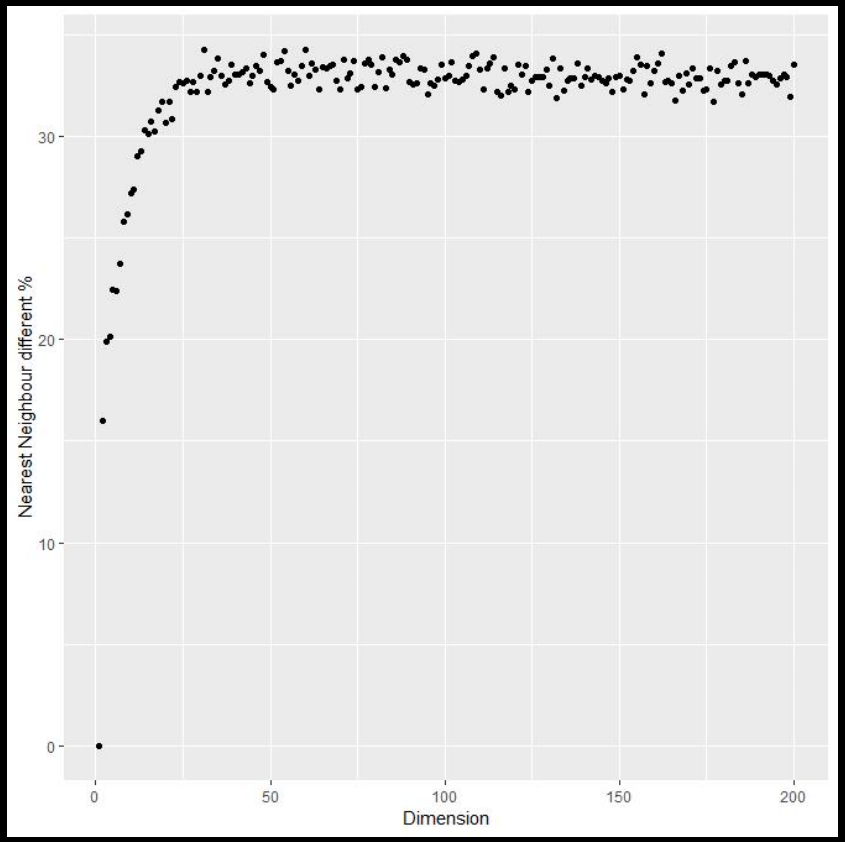
Observation id	min-max nn	mean-sd nn	median-iqr nn	median-mad nn
93	64	64	56	64
94	72	72	80	80

% change of nearest neighbours

Without outliers



With an outlier



Our experiment with real data

- 12000+ datasets
  - Around 200 sources, many variants
- 4 normalization methods
- 12 outlier detection methods
  - Performance = area under ROC curve

- How does normalization affect outlier detection?

Mixed effects models

- Model 1
- $perf \sim out + norm + (1|source)$

- Model 2
- $perf \sim out \cdot norm + (1|source)$

- out – outlier detection method
- $norm$ – normalization method
- $source$ – source dataset
- $perf$ – performance

- ANOVA prefers Model 2 with a p-value of 2.2×10^{-16}

Sensitivity to normalization

- ξ-Sensitivity to normalization
  - For a given outlier method
  - $(\max \text{perf} - \min \text{perf}) > \xi$

Filename	LOF (mean-sd)	LOF (median-iqr)	LOF (median-mad)	LOF (min-max)
Dataset 1 (abalone)	0.65	0.65	0.64	0.66
Dataset 2 (annthyroid)	0.70	0.86	0.90	0.56

Datasets sensitive to normalization

# of outlier methods	ξ=0.05	ξ=0.10	ξ=0.20
0	412	1977	5966
1	1002	2009	2330
2	656	1142	939
3	627	899	584
4	692	719	426
5	573	689	361
6	530	697	281
7	695	671	230
8	839	672	154
9	994	642	152
10	1205	532	111
11	1361	498	66
12	2029	468	15

$$norm\ sensitivity \sim\ f(dataset, outlier\ method)$$

# of outlier methods	ξ=0.05	ξ=0.10	ξ=0.20
0	412	1977	5966
1	1002	2009	2330
2	656	1142	939
3	627	899	584
4	692	719	426
5	573	689	361
6	530	697	281
7	695	671	230
8	839	672	154
9	994	642	152
10	1205	532	111
11	1361	498	66
12	2029	468	15

Datasets affected by a single outlier method

Outlier method	# datasets
COF	224
FAST ABOD	888
INFLO	122
KDEOS	585
KNN	31
KNNW	3
LDF	273
LDOF	66
LOF	46
LOOP	14
ODIN	71
SIMLOF	7
Total	**2330**

Takeaway

- Not to treat normalization as fixed
- Effect of normalization depends on datasets and outlier detection methods

https://github.com/sevvandi/Outliers-Norm-Instance

THANK YOU!