Similarity Based on Resample Exposure

Anton D. Lautrup, Hafiz Saud Arshad, Tobias Hyrup,
Muhammad Rajabinasab, Arthur Zimek, and Peter Schneider-Kamp
University of Southern Denmark, Odense, Denmark

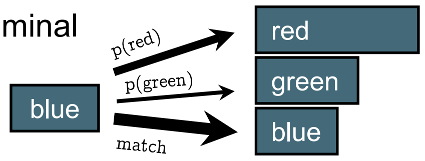
Problem: Existing measures of proximity in heterogeneous data discards most data insights beyond mere matching and therefore causes unstable results due to no granularity.

Goal: Develop a data-driven metric that accounts for frequency of categorical levels and employ distribution shape of the numerical variables.

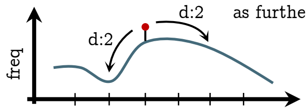
Solution: Resample exposure (REX) similarity

$$\mathrm{REX}(q,t) = \sum_{k}^{m} \begin{cases} 0 & \text{if } q_k, \text{or } t_k \text{ is missing} \\ 1 & \text{if } q_k = t_k \\ \Pr(t_k) & \text{if } k\text{th variable is nominal} \\ \left(1 - \dfrac{|q_k - t_k|}{\mathrm{rng}_k}\right)\left(1 - \dfrac{b(q_k t_k)}{\mathrm{trav}_k}\right) & \text{if } k\text{th variable is quantitative} \end{cases}$$

Nominal

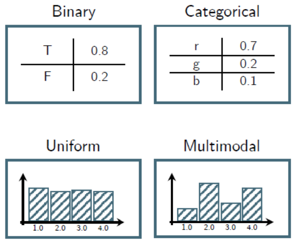


Quantitative



Close values that are downhill are counted as further away

	Binary q	Binary t	GOW	REX	Categorical q	Categorical t	GOW	REX
A	T	T	1	1	r	r	1	1
B	T	T	1	1	g	r	0	0.7
C	F	F	1	1	b	r	0	0.7
D	F	T	0	0.8	r	g	0	0.2
E	T	F	0	0.2	b	g	0	0.2
F	T	F	0	0.2	r	b	0	0.1



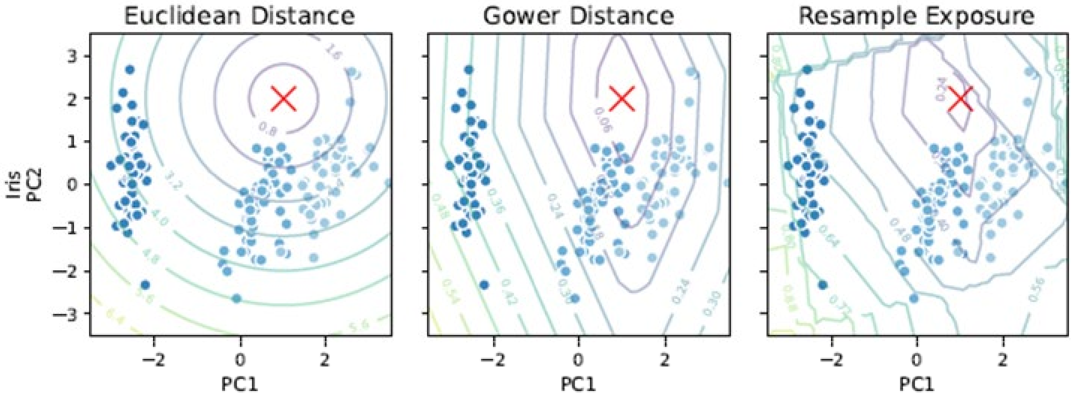
T	0.8
F	0.2

r	0.7
g	0.2
b	0.1

	Uniform q	Uniform t	GOW	REX	Multimodal q	Multimodal t	GOW	REX
A	1.0	1.0	1	1	1.0	1.0	1	1
B	2.0	1.0	0.66	0.66	1.0	2.0	0.66	0.66
C	1.0	2.0	0.66	0.66	2.0	1.0	0.66	0.5
D	1.0	3.0	0.33	0.33	1.0	3.0	0.33	0.22
E	2.0	4.0	0.33	0.33	2.0	4.0	0.33	0.22
F	1.0	4.0	0.0	0.0	4.0	1.0	0.0	0.0

Overall Score

	GOW	REX
A	4	4
B	2.33	3.03
C	2.33	2.86
D	0.66	1.55
E	0.66	0.95
F	0.0	0.3


Euclidean Distance
Gower Distance
Resample Exposure
Iris
PC2
PC1
PC1
PC1