8th Conference on the Use of R in official Statistics (uRos2020)
Wien, 2-4 December 2020

DISTANCES WITH MIXED-TYPE VARIABLES: SOME PROPOSALS

Marcello D'Orazio
marcello.dorazio(at)istat.it
marcello.dorazio(at)fao.org

Italian National Institute of Statistics – Istat, Rome, Italy
Office of Chief Statistician, Food and Agriculture Organization (FAO) of the UN, Rome, Italy

Nearest Neighbor methods are very popular in statistics:

- Clustering observations (unsupervised learning)
- Supervised learning: regression or classification
- Official statistics: donor-based imputation, data fusion, record linkage, etc.

The NN methods rely on calculation of distance between observations

Many distance functions when ALL the variables are of the same type (all categorical or all quantitative)

Few options with mixed-type variables (mix of categorical and quantitative)
Most popular: **Gower's distance**

Gower's distance: complement of the *Gower's similarity coefficient*

$$d_{G,ij} = 1 - s_{G,ij} = \frac{\sum_{t=1}^{p} \delta_{ijt}\, d_{ijt}\, w_t}{\sum_{t=1}^{p} \delta_{ijt}\, w_t}$$

Weighted average of distances calculated variable by variable

w_t is the weight of t-th variable

Equal weights ($w_t = 1$) give the unweighted ("standard") version

Type of variable	d_{ijt}	δ_{ijt}	Note
binary symmetric	0 if $x_{it} = x_{jt}$ 1 if $x_{it} \neq x_{jt}$ 1 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if x_{it} or x_{jt} or both are missing	Corresponds to $1 -$ *simple matching coefficient*
binary asymmetric	0 if $x_{it} = x_{jt} = 1$ 1 otherwise 1 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if $x_{it} = x_{jt} = 0$ 0 if x_{it} or x_{jt} or both are missing	corresponds to the $1 -$ *Jaccard index*
categorical nominal (more than two categories)	0 if $x_{it} = x_{jt}$ 1 if $x_{it} \neq x_{jt}$ 1 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if x_{it} or x_{jt} or both are missing	Corresponds to the $1 -$ *Dice* Applied to the dummies associated to the original variable $1-$simple matching on the starting categorical variable
measured on an interval or ratio scale	$\lvert x_{it} - x_{jt} \rvert / R_t$ 1 if x_{it} or x_{jt} or both are missing	1 if both the variables are nonmissing 0 if x_{it} or x_{jt} or both are missing	$R_t = \max(x_t) - \min(x_t)$ is the range of the kth variable d_{ijt} is *the Manhattan* or *city-block distance* scaled by the range

Kaufman and Rousseeuw (1990) and Podani (1999) different proposals for handling **categorical ordered variables**

Gower's distance is an average of the distances d_{ijt} :

$$0 \leq d_{ijt} \leq 1 \text{ for each variable} \quad \Rightarrow \quad 0 \leq d_{G,ij} \leq 1$$

a variable with a missing value does NOT contribute to the average

Several implementations

`daisy` function in **cluster** (Maechler et al, 2019)

`gowdis` function in **FD** (Laliberté et al, 2014) implements also the various options for calculating distance on categorical ordered variables.

- package **gower** (van der Loo, 2020): Gower's distance as well as the top-*n* matches between records; it is very efficient and fast
- package **kmed** (Budiaji, 2019): Gower's and other distance functions for mixed-type variables (not categorical ordered)
- package **proxy** (Meyer, 2020): Gower's distance and many other ones

Gower (1971): "the decision on a rational set of weights is difficult"

Discussion often misleading, because in the unweighted Gower's distance the variables have an **unbalanced contribution** to the overall distance

Example of unbalanced contribution

Obs.	Sex	Age
1	F	18
2	F	78
3	M	20

$$d_{G,12} = \frac{1}{2} \times 0 + \frac{1}{2} \times \frac{|18 - 78|}{100} = 0.30$$

$$d_{G,13} = \frac{1}{2} \times 1 + \frac{1}{2} \times \frac{|18 - 20|}{100} = 0.51$$

Units with different gender are **farther** than units with the same gender but showing a huge distance in terms of age

Distance on categorical variables is **0** or **1**

Distance on variables measured on interval/ratio scale: 0 only when $x_{it} = x_{jt}$

1 only when $|x_{it} - x_{jt}| = R_t$, rare if there are outliers!!!!

Estimation of the range (R_t) is highly affected by outliers

Change the way of calculating the distance on interval/ratio-scaled variables in the Gower's formula to:

1) reduce impact of outliers
2) balance the contribution of variables of different type

Problem (1): **scaling by IQR** (Inter Quartile Range)

$$d_{ijt} = \begin{cases} \dfrac{|x_{it} - x_{jt}|}{IQR_t}, & \text{if } |x_{it} - x_{jt}| < IQR_t \\ 1, & \text{otherwise} \end{cases}$$

Problem (2a): modifications based on kernel density estimation

$$d_{ijt}^{(kde)} = \begin{cases} 0, & \text{if } |x_{it} - x_{jt}| \leq h_t \\ \dfrac{|x_{it} - x_{jt}|}{g_t}, & \text{if } h_t < |x_{it} - x_{jt}| < g_t \\ 1, & |x_{it} - x_{jt}| \geq g_t \end{cases}$$

h_t the *window width* (*bandwidth* in the kernel density estimation)

$$h_t = c\frac{1}{n^{1/5}} min\left\{s_t, \frac{IQR_t}{1.34}\right\}$$

s_t is the estimated standard deviation for the t-th variable

$c = 1.06$ or $c = 0.9$ (Silverman, 1986, p. 48)

g_t can be the range or the IQR

Problem (2b): modifications based on *k*-NN

$$d_{ijt}^{(knn)} = \begin{cases} 0, & \text{if } x_{jt} \text{ is one of the } k \text{ nearest neighbors of } x_{it} \\ \dfrac{|x_{it} - x_{jt}|}{g_t}, & x_{jt} \text{ not one of } k \text{ near. neigh. of } x_{it} \; AND \; |x_{it} - x_{jt}| < g_t \\ 1, & \text{if } |x_{it} - x_{jt}| \geq g_t \end{cases}$$

$k = \sqrt{n}$ (well-known rule of thumb)

Step 1) n = 500 obs. generated from multivGaussian = 100, σ = 20, and correlation matrix

	X_1	X_2	X_3	X_4	X_5
Y	0.8	0.4	0.8	0.4	0.5
X_1		0.2	0.4	0.2	0.3
X_2			0.2	0.2	0.3
X_3				0.2	0.2
X_4					0.2

Step 2) categorization of variables (X_2, … , X_5) **OR** (X_3, … , X_5)

Step 3) WithOut outliers in X_1 **OR** with outliers in X_1 (approx. 2% obs.)

Step 4) 33% of values of Y are randomly deleted

Step 5) missing Y values are **imputed** with Nearest Neighbor donor hotdeck, distance calculated using standard and modified Gower's dissimilarity

1,000 runs for each combination

Simulation results

Case	Eval. criterion	Scaled by range				Scaled by IQR			
		no.mod	kde1	kde2	knn	tr.IQR	IQR.kde1	IQR.kde2	IQR.knn
1 cont. 4 cat. No outl.	sB	-0.0645	-0.0636	-0.0771	-0.0717	-0.0605	-0.0610	-0.0615	-0.0655
	sAbsB	6.6765	6.7514	6.7181	6.6951	6.7961	6.6521	6.6343	6.5968
	sBmean	0.0601	0.0597	0.0637	0.0623	0.0591	0.0593	0.0592	0.0604
	sDiff.qqs	-0.0174	-0.0167	-0.0223	-0.0191	-0.0160	-0.0181	-0.0187	-0.0183
1 cont. 4 cat. With outl.	sB	0.1713	0.1610	0.1679	0.1794	-0.0203	-0.0176	-0.0214	0.1195
	sAbsB	6.8998	6.9732	6.9522	6.9180	6.9420	6.8274	6.8022	6.8474
	sBmean	0.0569	0.0542	0.0558	0.0595	0.0008	0.0004	0.0005	0.0414
	sDiff.qqs	0.0501	0.0481	0.0498	0.0526	-0.0061	-0.0037	-0.0049	0.0335
2 cont. 3 cat. No outl.	sB	-0.0019	0.0066	-0.0007	0.0065	-0.0199	-0.0070	-0.0202	-0.0046
	sAbsB	6.1527	6.1819	6.1831	6.1677	6.3745	6.4221	6.4322	6.3441
	sBmean	0.0096	0.0118	0.0097	0.0119	0.0041	0.0079	0.0037	0.0085
	sDiff.qqs	-0.0022	0.0011	-0.0016	0.0009	-0.0069	-0.0013	-0.0061	-0.0012
2 cont. 3 cat. With outl.	sB	0.2014	0.1896	0.1946	0.2357	0.0271	0.0144	0.0308	0.2030
	sAbsB	6.7073	6.7749	6.7769	6.7696	6.4768	6.5332	6.5344	6.5547
	sBmean	0.0649	0.0617	0.0630	0.0752	0.0128	0.0095	0.0144	0.0659
	sDiff.qqs	0.0598	0.0554	0.0568	0.0692	0.0121	0.0082	0.0109	0.0618

- ✓ In presence of outliers in X_1, scaling by IQR tend to perform better than by scaling by the range; in particular in preserving the true distribution

- ✓ The modification based on the kernel density estimation with c =1.06 (“kde1” in Table) seems to perform slightly better for all the assessment criteria with the exception of the average absolute difference between imputed and true values (“sAbsB”)

- ✓ The results are very close and no one of the proposed modifications outperforms the other

The modifications of the standard unweighted Gower distance for calculating the distance on ratio-scaled variables go in the desired direction

Further investigation is needed

- when the quantitative variables show a skewed distribution
- in other problems: clustering, *k*-NN classification, etc.

Have to be implemented in R (package **StatMatch**)

Warning: **parsimony in selecting the variables** on which to calculate the distance, also to avoid the curse of dimensionality:

"everything starts being close to everything"

Thank you for your attention