

The following is a scheme for weighting variables in a mixed-type dataset to convert into a Gower matrix.

x is a variable of dataset X .

If x is a numeric variable, then $w \leftarrow 1$.

Otherwise, if x is a categorical variable, then:

$$w \leftarrow \left[\sum_k^K \widehat{\text{Var}}(h_k) \right] \left[1 - \sum_k^K \widehat{\text{Var}}(h_k) \right] \left[\widehat{\text{Kurtosis}}(H) \right] \quad (1)$$

H is the one-hot encoding of x and h_k is the column of H corresponding to the k^{th} class of x .

The first term is zero when all the values of the variable are the same, whereas the second term is zero when they're all different. The third term increases as the number of unique values increases.