Metrics for Mining Multisets

Tim K. Cocx *tcocx@liacs.nl*

Walter A. Kosters *kosters@liacs.nl*

Jeroen F. J. Laros *jlaros@liacs.nl*

Universiteit Leiden

Metrics for multisets

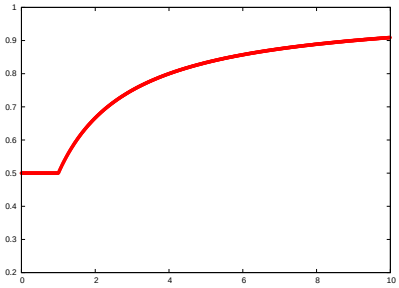
We propose a new class of distance measures (metrics) designed for multisets, both of which are a recurrent theme in many data mining applications. One particular instance of this class originated from the necessity for a clustering of criminal behaviours. The class generalizes Canberra and Jaccard distance.

Different functions

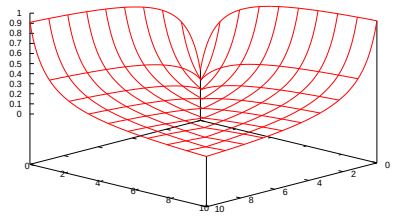
For a multiset X, let $S(X)$ denote its underlying set. For multisets $X, Y \subseteq \{1, 2, \ldots, n\}$ we define $d_f(\emptyset, \emptyset) = 0$ and

$$d_f(X,Y) = \frac{\sum_{i=1}^{n} f(x_i, y_i)}{|S(X) \cup S(Y)|}$$

if both X and Y are non-empty. Here x_i (resp. y_i) is the number of times that i occurs in X (resp. Y). These distance measures are parameterized by the function f which, given a few restrictions, will always produce a valid metric. This flexibility allows these measures to be tailored for many domain-specific applications. An important restriction on the function is that given the supremum M of f, $f(0, x) \geq M/2$ for all $x > 0$.



One way to make such a function f is by using a suitable function g of one variable, composing the function $f(x, y) = |g(x) - g(y)|$. A function g that can be used like that is in the picture above: $g(x) = x/(x+1)$ if $x > 1/2$ and $1/2$ otherwise. After the composition, the function f_1 will look like this:



Applications

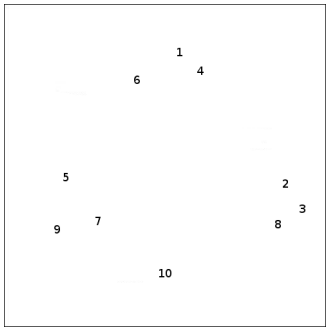
We use an example database with ten criminals and four crime categories, and the numbers of crimes committed. E.g., the multiset for criminal 8 has 2 crimes of type $\mathcal{A}$ and 4 crimes of type $\mathcal{B}$.

	1	2	3	4	5	6	7	8	9	10
$\mathcal{A}$	0	2	10	0	0	0	0	2	0	2
$\mathcal{B}$	0	0	0	2	0	0	2	4	0	2
$\mathcal{C}$	0	0	0	0	1	0	2	0	3	2
$\mathcal{D}$	0	0	0	0	0	1	1	0	5	2

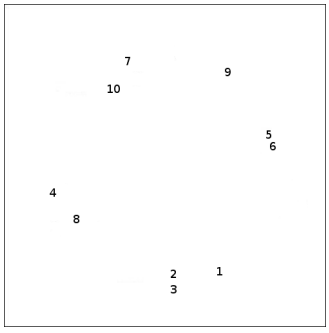
Category $\mathcal{A}$ consists of relatively light crimes, and the severity increases as we go to $\mathcal{D}$. Note that "criminal" 1 is innocent.

We first show a clustering with the function

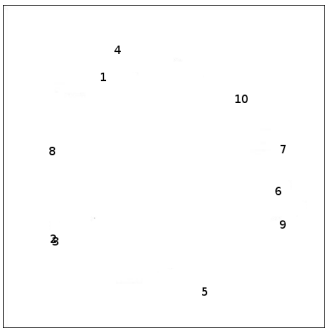
$$f_1(x,y) = \frac{|x-y|}{(x+1)(y+1)}$$



Next we use the same f_1 but with weights 1, 10, 100 and 1000. We now see that criminals 7 and 10 are very close together, but at the same time, criminals 2 and 3 also stay close. "Criminal" 1 is surprisingly rather close to the two criminals who have committed relatively light crimes. The reason that criminals 5 and 6 are close together is because they are one-time offenders, and have a large distance to the rest of the group. Clearly, one must be aware of the danger that clusterings can be easily affected by changing the metric.



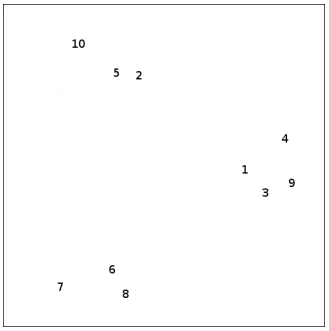
Here we see a clustering where we consider the multisets as normal sets, so the number of crimes is not taken into account. Notice that criminals 2 and 3 now have distance zero to each other.



Finally we use a totally different formula:

$$f_2(x,y) = \frac{3}{2} - f_1(x,y)$$

for $(x, y) \neq (0, 0)$, where f_1 is the function described above, and $f_2(0, 0) = 0$. This results in a *dissimilarity* measure, rather than a *similarity* measure.



The DALE project is financed in the ToKeN program of the Netherlands Organisation for Scientific Research (NWO) under grant number 634.000.430.

Data Assistance for Law Enforcement

http://www.dale.liacs.nl/

NWO ToKeN2000