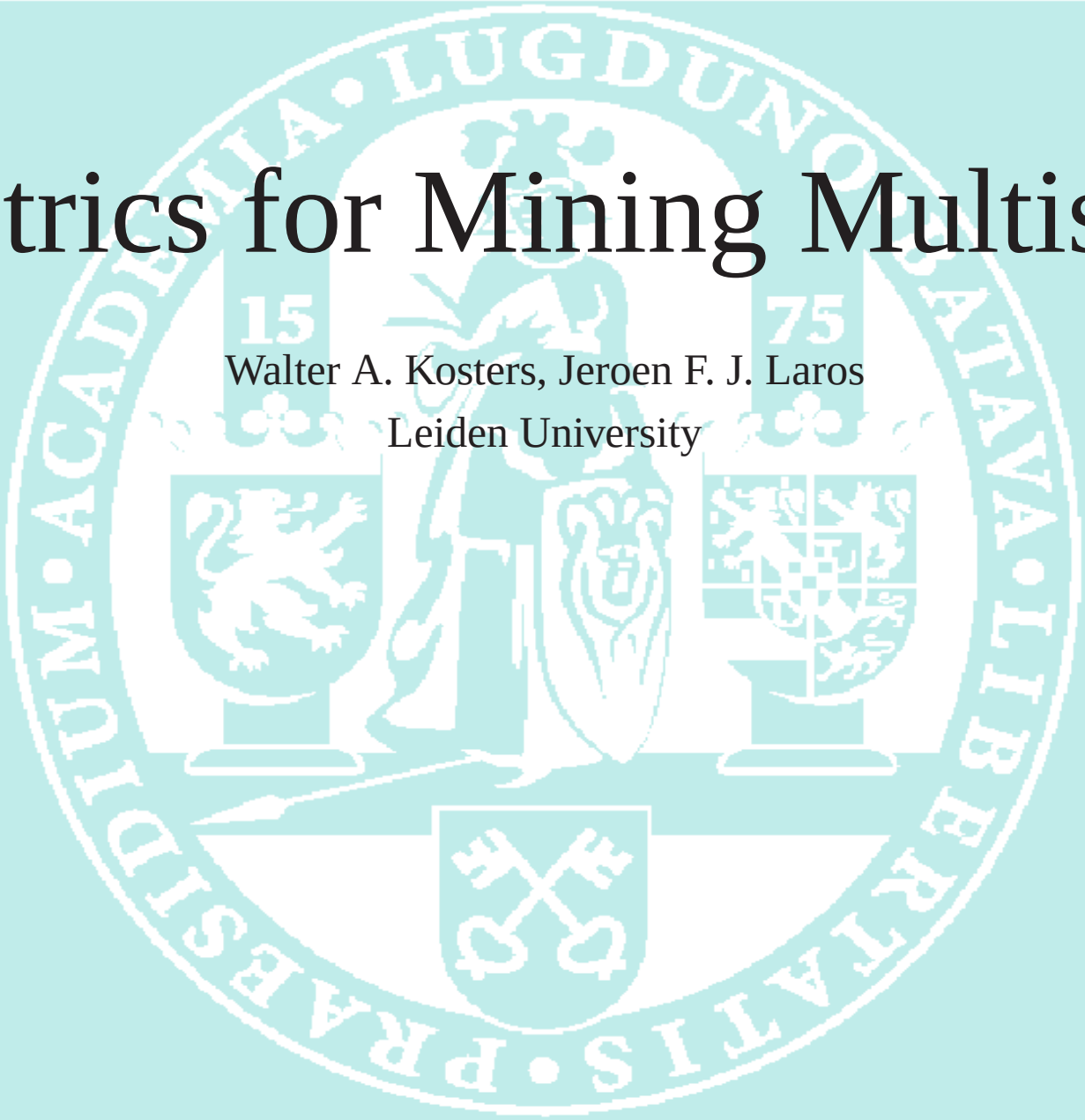
Metrics for Mining Multisets

Walter A. Kosters, Jeroen F. J. Laros
Leiden University



Multisets

A **multiset** (or "bag") is a set where not all elements have to be distinct.

Multiset X and the underlying set $S(X)$.

A vase with 4 blue marbles and 2 red marbles is a multiset X. The **underlying set** ($S(X)$) is $\{\text{blue}, \text{red}\}$.

In practice many things are multisets: all strings of DNA of length 20 in the human genome, criminal records, the words in a document ("bag of words"), . . .

We propose a new generic **distance measure**:

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

Here x_i is the number of elements in category i, for example the number of blue elements in multiset X.

We divide by the number of categories in X and Y.

The function f with finite supremum M has four restrictions:

- Symmetry: $f(x,y) = f(y,x)$.
- Distance to itself: $f(x,x) = 0$.
- Distance to the empty set: $f(x,0) \geq M/2$.
- Triangle inequality: $f(x,y) \leq f(x,z) + f(z,y)$.

The parameter f can be tuned. It specifies the difference between the number of occurrences of a particular element in two multisets. Constructing such a function is natural and can easily be done by domain experts.

Q: What is the difference between 2 and 10 blue elements?

A: $f(2, 10)$.

Ten criminals, four crimes

	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

Here we see a dataset with crimes $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ of increasing severity, and criminals ranging from 1 to 10. For each criminal the number of crimes in each category is given.

- 1: Innocent.
- 2: Incidental minor criminal.
- 6: One-time offender of a serious crime.
- 9: Severe criminal.
- 10: All-rounder.

We use a **push and pull** algorithm to cluster the data once all pairwise distances are determined.

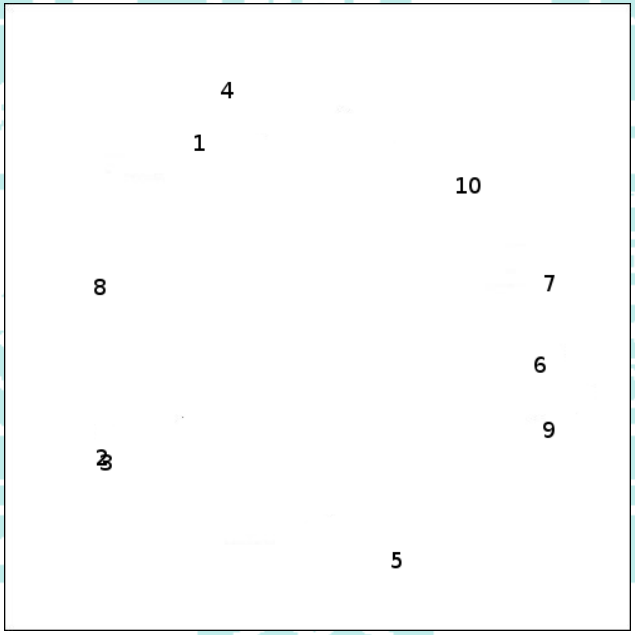
This type of multi dimensional scaling algorithms iterates over all pairs of (randomly initialized) points and corrects the distances until an optimum is reached.

The well-known **Jaccard** distance measure is a function defined on sets.
This measure is used in market basket analysis.

$$d(X,Y) = \frac{|X \backslash Y| + |Y \backslash X|}{|X \cup Y|}$$

By choosing the right f, we get this formula.

Visualization of the data with the Jaccard distance



Notice that criminal 2 and 3 are considered the same. They both are minor criminals, but the number of crimes differ greatly.

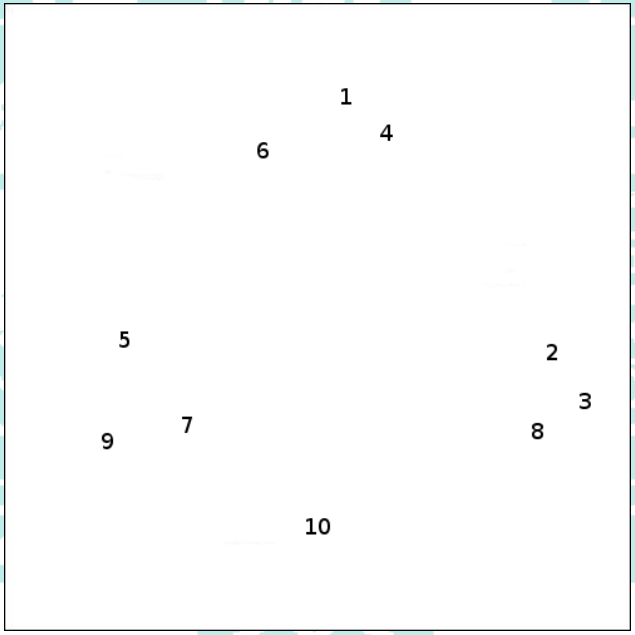
To distinguish between the number of crimes, we need a more complicated formula.

$$f_0(x,y) = \frac{|x-y|}{(x+1)(y+1)} \text{ or } f_1(x,y) = \frac{|x-y|}{x+y+1}$$

Notice that the difference between 0 and 1 is large, but the difference between 9 and 10 is not.

This comes in handy when comparing two criminals, the difference between an innocent man and someone who had committed a crime is larger than two criminals that differ in the number of crimes they have committed.

Visualization of the data with the multiset distance



Here we see a clustering of the ten criminals, but now with our f_0 function.

Notice that criminal number 2 and 3 still are close together, but do have a considerable distance.

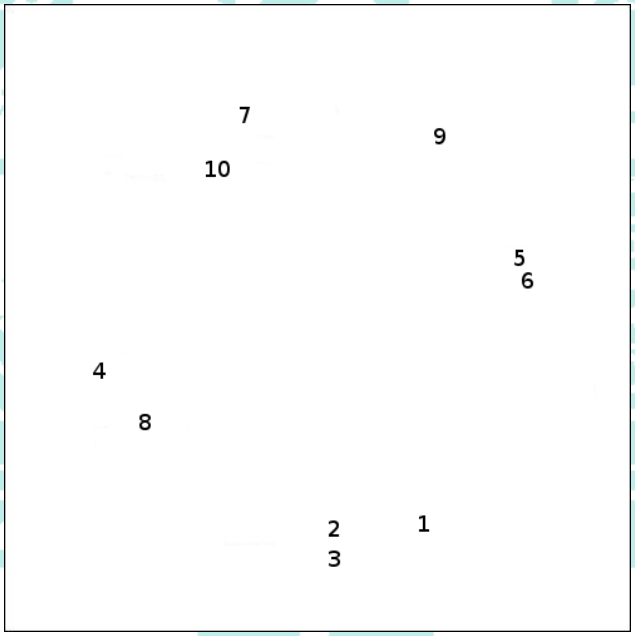
Adding weights

Weights can be implemented by multiplying the number of elements of a category with the specified weight.

Example: $\{\text{red}, \text{red}, \text{blue}, \text{blue}, \text{blue}\}$ with weight 2 for red and weight 1 for blue, results in $\{\text{red}, \text{red}, \text{red}, \text{red}, \text{blue}, \text{blue}, \text{blue}\}$.

Now we add weights $(1, 10, 100, 1000)$ to the categories of crimes in our example. This weight indicates the severity of the crime.

Visualization of the data with weighted multiset distance



Notice that criminal 2 and 3 now are close to number 1 (the innocent one). Also note that criminal 7 and 10 are now very close together, these are convicted of heavy crimes.

Other applications

Other applications include, but are not limited to:

- Genetic research, we mention:
  - Comparison of DNA using small fragments of a fixed length.
  - Comparison on the occurrence of SNPs.

  Both techniques can be used for phylogenetic tree construction and crime investigation.
- Comparison of documents: plagiarism, textmining, ...

Conclusions and further research

We have devised an interesting class of distance measures for multisets.

For the clustering of criminal behaviour through time, we want to make a string of multisets. Then we want to use this distance measure to align these strings. Doing so will hopefully give more insight in criminal behaviour through time. The method might also be used to make predictions about starting criminals.

Questions

Universiteit Leiden