Alignment of Multiset Sequences

Tim K. Cocx
tcocx@liacs.nl

Walter A. Kosters
kosters@liacs.nl

Jeroen F. J. Laros
jlaros@liacs.nl

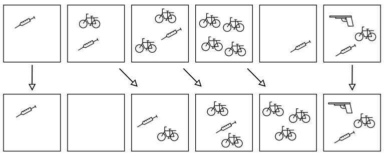
Universiteit Leiden

Multiset Sequence Alignment

We introduce a new metric for the similarity between sequences of multisets [1]. This distance is based on a well-defined distance measure for multisets. Various types of alignment are used to find different aspects of similarity in two sequences. Applications of this metric to the analysis of criminal careers are reviewed.

Various visualisations

The edit distance is a distance measure for sequences. It gives a penalty for two different elements and for an insertion or a deletion of an element. By calculating this distance, we implicitly calculate an optimal alignment.



As seen in the picture above, an alignment is an ordering of the elements in such a way that the edit distance is minimal. The order of the elements however, must be preserved.

There are a number of parameters to this approach, using local or global alignment (so compare whole careers or to detect sub-careers). We can expand careers to include periods of inactivity or we can omit that to compare criminals based on their actions only. Furthermore we can use several methods to scale the distances between 0 and 1 by either dividing by the maximum length in each pairwise comparison, or by dividing by a large number (the length of the largest sequence in the database). The former will emphasise pairwise similarities, the latter will give a global picture of the structure of the database.

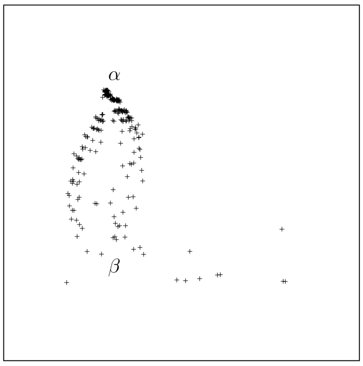
Applications

Our target database is the HKA database of the Dutch National Police (KLPD). This database consists of all recorded criminals and their crimes committed per year. Each year can therefore be seen as a multiset of crimes committed by that person. The whole sequence of these multisets is a criminal career.

	1999	2000	2001	2002	2003
A	$\{1,2\}$	$\{3\}$	$\{1,1,3\}$	$\{2,3\}$	$\{3\}$
B	$\{3,3\}$	$\emptyset$	$\{3,4\}$	$\{3,3\}$	$\{3,4\}$

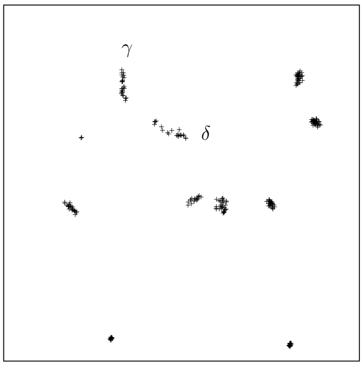
Example database: Criminal A and B and the categories in which their committed crimes are categorised.

Using these approaches we can make several views of the database.



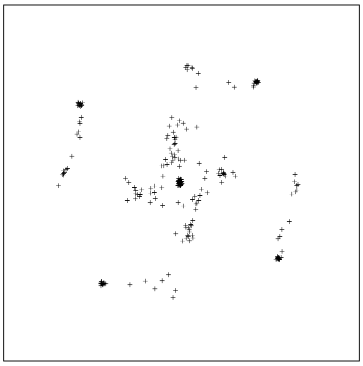
Database of criminal careers. Global alignment, absolute scaling, non-expanded careers.

We see a large cluster denoted by α, these are all short careers. As we move from α to β the length of the careers increases.

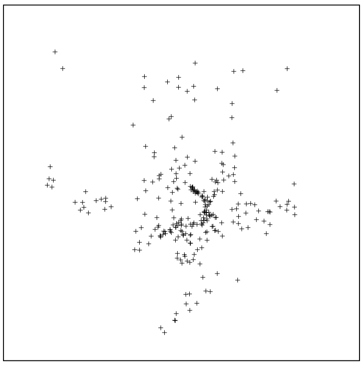


Database of criminal careers. Global alignment, relative scaling, non-expanded careers.

The clusters indicated by γ and δ contain only long careers. The other clusters have careers of roughly the same size and their elements have been clustered according to the similarity in career.



Database of criminal careers. Local alignment, absolute scaling, non-expanded careers.



Database of criminal careers. Local alignment, relative scaling, expanded careers.

In the last two pictures, we see something different emerging: the long careers are no longer put into the same cluster, but more emphasis is given to similarity in the nature of criminal activity. For example, the central cluster in the third picture consists mainly of criminals who have committed non-violent crimes involving large sums of money.

Further research

We want to use this new metric to improve temporal extrapolation for criminal careers [2].

References

[1] Kosters, W.A., and Laros, J.F.J., Metrics for mining multisets, Research and Development in Intelligent Systems XXIV, Proceedings of AI-2007, the Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence (SGAI 2007), Springer, pp. 293–303, 2007.

[2] Cocx, T.K., Kosters, W.A., and Laros, J.F.J., Temporal extrapolation within a static clustering, Proceedings of ISMIS 2008, Springer, LNAI 4994, pp. 189–195, 2008.

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

DALE
Data Assistance for Law Enforcement

http://dale.liacs.nl/

NWO ToKeN2000