Metrics for Mining Multisets *

Walter A. Kosters Jeroen F. J. Laros

Leiden Institute of Advanced Computer Science, Universiteit Leiden, The Netherlands

1 Introduction

A *multiset* (also referred to as a bag) is a set (collection of elements where the order is of no importance), where the elements do not need to be unique. A vase with n blue and m red marbles is a multiset for example.

We propose a new class of distance measures (metrics) designed for multisets, both of which are a recurrent theme in many *data mining* [2] applications. One particular instance of this class originated from the necessity for a clustering of criminal behaviours. Here the multisets are the crimes committed in one year. This metric generalises well-known distance measures like the Jaccard and the Canberra distance.

These distance measures are parameterised by a function f which, given a few simple restrictions, will always produce a valid metric. This flexibility allows these measures to be tailored for many domain-specific applications. The metrics in this class can be efficiently calculated. In the full paper, all proofs are given and various applications are shown.

2 The Metric

In order to produce a decent distance measure d_f, we carefully choose a function $f(x, y)$ that denotes the difference between the number of elements x and y of a specific type. This can not be any function; it has to have a finite supremum M and $f(x, 0)$ must be larger than or equal to $M/2$ (for $x > 0$) in order for the triangle inequality to hold. The function should also be symmetric and $f(x, x)$ should be zero. Also, the triangle inequality must hold for f itself as well. With this f we can now define a metric for multisets. We consider multisets X, Y over $\{1, 2, \ldots, n\}$, and let $x_i \in \mathbb{Z}_{\geq 0}$ (resp. y_i) be the number of times that i ($i = 1, 2, \ldots, n$) occurs in X (resp. Y). For a multiset X, let $S(X)$ denote its underlying set. We define $d_f(\emptyset, \emptyset) = 0$ and for multisets X, Y:

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

if X or Y is non-empty.

The application of weights for certain elements can be done by multiplying the appropriate number of elements by the weight. An important characteristic of these metrics is that the distance increases significantly when we add an extra dimension. This is not the case in other well-known metrics like the standard Euclidean distance.

3 Example

We use the following function to give an impression of the measure. As the expert defined function, we take $f(x, y) = |x - y|/(x + 1)(y + 1)$.

As a test case, we made the following synthetic dataset with fictional crimes $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ of increasing severity, and criminals ranging from 1 to 10 as seen in Table 1. For each criminal the number of crimes in each category is given. For instance, 1 is innocent, 2 is an incidental small criminal, 6 is a one-time offender of a serious crime, and 10 is a severe criminal.

*Published in: 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 (M. Bramer, F. Coenen, M. Petridis, editors), Springer, pp. 293–303, Cambridge, UK, 2007.

This research is part of the DALE (Data Assistance for Law Enforcement) project as financed in the ToKeN program from the Netherlands Organization for Scientific Research (NWO) under grant number 634.000.430.

	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

Table 1: Ten criminals $(1, 2, \ldots, 10)$, four crimes $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$

We use a distance preserving dimension reduction algorithm described in [1] to obtain the visualisations in Figure 1. When we choose the same weights for all categories, we obtain the picture on the left. Number 2 and 3 are close together, and 8 is near there too. This is what we would expect. Number 9, 7 and 5 are close together too, as could be expected; 1, 4 and 6 however are close to each other because they have a large distance to all others.

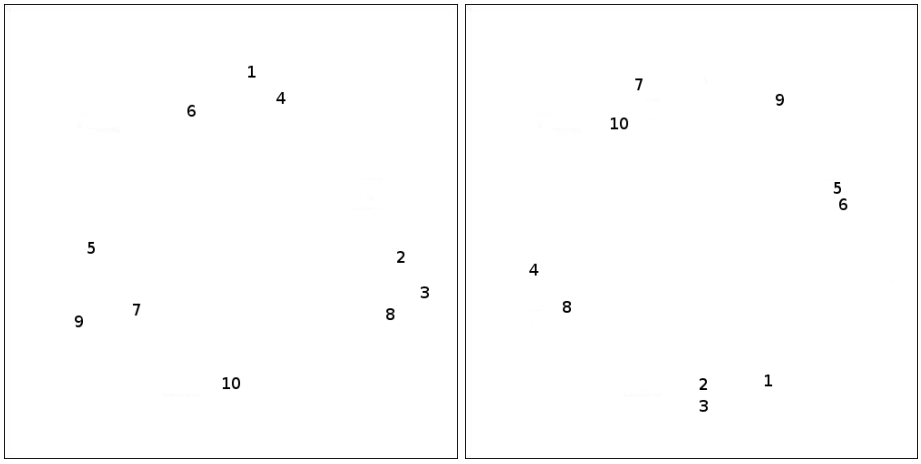

Figure 1: Two different clusterings for ten criminals

If we apply weights $(1, 10, 100, 1000$ for $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$ respectively) to accentuate the severity of a crime, we obtain the picture on the right. This gives more insight into the nature of the criminal, 10 and 7 are close together for example (both heavy criminals), while 2, 3 and 1 are close to each other because they have committed no or relatively light crimes.

4 Conclusions and Further Research

We have proposed a flexible distance measure, that is suitable in many fields of interest. It can be fine tuned to a large extent. This may result in different visualisations, illustrating different aspects of the data (see Figure 1).

We can use this measure as a basis for further analysis, like the analysis of criminal careers. In that case, we suggest that the distance measure is used as a basis for *alignment* to make the best match between two careers. By doing this, and by comparing sub-careers, we might be able to extrapolate criminal behaviour based upon the criminal record through time.

References

[1] Kosters, W.A., Wezel, M.C. van, Competitive neural networks for customer choice models, in E-Commerce and Intelligent Methods, volume 105 of Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2002, pp. 41–60.

[2] Tan, P.-N., Steinbach, M., Kumar, V., Introduction to data mining, Addison-Wesley, 2005.