# Metrics and visualisation for crime analysis and genomics

Laros, J.F.J.

# Chapter 6

# Metrics for Mining Multisets

In this chapter, 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.

These distance measures are parametrised 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.

In this chapter, the metrics are applied in bio-informatics (genomics), criminal behaviour clustering and text mining. The metric we propose also is a generalization of some known measures, e.g., the Jaccard distance and the Canberra distance. We discuss several options, and compare the behaviour of different instances.

## 6.1 Introduction

In many fields data mining is applied to find information in large amounts of data. A few example areas are bio-informatics, crime analysis and of course computer science itself. In data mining, *multisets* (also referred to as bags) are a recurring theme. Finding *distance measures* or *metrics* (for multisets) is one aspect of data mining [67]. When a suitable measure is found, many types of analysis, such as clustering, can be performed on specific documents, DNA and other instances of multisets.

The reasons for finding distance measures are very diverse. In crime analysis [9] for example, it is possible to determine the distance between two criminals based on their behaviour (their crime record). In bio-informatics comparing two species with only the information of their DNA (or short fragments of it) can be done. This is especially useful in forensic applications where DNA strands are frequently damaged, so the fragments that are extracted from samples cannot

be given a place on the genome. Even without the information of the placement of the DNA fragments found, it is possible to differentiate between species and even individuals by using techniques described in this chapter. We finally mention market basket analysis, where distances between multisets are basic for further analysis. As a motivating example, the distance between two customers can or cannot take into account the numbers of purchases of individual products (thus providing either multisets or sets), and it is also possible to stress the difference between 1 and 2 sales on the one hand and, e.g., 41 and 42 sales on the other hand.

In Section 6.3 we give a new class of distance measures that are suitable for comparing multisets. The class has a parameter $f$ (a function) that has a couple of simple properties which, if met, will always produce a valid metric. To the best of our knowledge, the class is new, and generalizes several of the more well-known distance measures mentioned in Section 6.2 and Section 6.3.

For different domains, different problems arise and different distance measures will be needed. For many of them, a tailor made function $f$ can be provided and if the given restrictions apply to $f$, no further effort has to be made with respect to the validity of the metric. We mention several examples in the applications in Section 6.4. Different choices of $f$ may lead to different visualisations. Furthermore, choosing such a function $f$ is rather straightforward and intuitively easier to do than constructing a metric directly.

## 6.2 Background

Finite multisets from a universe with $n$ elements can be viewed as points in $n$-dimensional space. For example, the multiset $\{a, b, a, a, b, a\}$ can be abbreviated to $\{a^4, b^2\}$ (since the order of elements is irrelevant) and by leaving out the element names, we get the vector $(4, 2)$ in 2-dimensional space. Several known distance measures can be applied. We mention the most important ones. In Section 6.3 we will show the relation with our metric. In all cases, we consider multisets $X, Y$ over $\{1, 2, \ldots, n\}$, and let $x_i \in \mathbb{R}_{\geq 0}$ (resp. $y_i$) be the number of times that $i$ ($i = 1, 2, \ldots, n$) occurs in $X$ (resp. $Y$).

- Minkowski distance of order $p$ [67]

$$d(X, Y) = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

  For $p = 1$, we get the Manhattan distance; for $p = 2$, we get the well-known Euclidean distance; if we let $p = \infty$, we get the Chebyshev distance or $L_\infty$ metric.

- Canberra distance

$$d(X, Y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{x_i + y_i}$$

When both $x_i$ and $y_i$ are zero, the fraction is defined as zero. Often the distance is divided by the number of indices $i$ for which at least one of $x_i$ or $y_i$ is non zero.

- Jaccard distance for sets [34]

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \;/\; \Big(n - \sum_{i=1}^{n}(1-x_i)(1-y_i)\Big)$$

- Bray-Curtis (Sorensen) distance (often used in botany, ecology and environmental science) [6]

$$d(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \;/\; \sum_{i=1}^{n}(x_i + y_i)$$

- Mahalanobis distance (generalized form of the Euclidean distance) [48] This is an example of a metric that requires a more complicated scheme: the covariance matrix of the data must be computed, which is quite time-consuming. We will not further discuss this type of metric here.

## 6.3 The metric

In this section we will define our new *class of metrics*. As a parameter we have a function $f$ that must meet several properties.

Let $f$ be a function $f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with finite supremum $M$ and the following properties:

$$\begin{aligned}
f(x,y) &= f(y,x) && \text{for all } x,y \in \mathbb{R}_{\geq 0} && (6.1)\\
f(x,x) &= 0 && \text{for all } x \in \mathbb{R}_{\geq 0} && (6.2)\\
f(x,0) &\geq M/2 && \text{for all } x \in \mathbb{R}_{>0} && (6.3)\\
f(x,y) &\leq f(x,z) + f(z,y) && \text{for all } x,y,z \in \mathbb{R}_{\geq 0} && (6.4)
\end{aligned}$$

For a multiset $X$, let $S(X)$ denote its underlying set. For multisets $X, Y$ with $S(X), S(Y) \subseteq \{1, 2, \ldots, n\}$ 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. Again, $x_i \in \mathbb{R}_{\geq 0}$ (resp. $y_i$) is the number of times that $i$ ($i = 1, 2, \ldots, n$) occurs in $X$ (resp. $Y$; usually $x_i$ and $y_i$ are integers); $|S(X) \cup S(Y)|$ is the number of elements in $X \cup Y$, seen as set. Note that $0 \leq d_f(X,Y) \leq M$, $d_f(X,Y) = d_f(Y,X)$ and $d_f(X,Y) = 0 \Rightarrow S(X) = S(Y)$. If $f$ also satisfies

$$f(x,y) = 0 \Rightarrow x = y \quad \text{for all } x,y \in \mathbb{R}_{\geq 0} \qquad (6.5)$$

we have $d_f(X,Y) = 0 \Rightarrow X = Y$. It is clear that properties (1), (2) and (4) must hold in order to ensure that we have a metric; indeed, just consider the case where $n = 1$.

The function $f$ 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. Also note that the function $f$ is defined for all positive real numbers; this property is only used when weights are involved (see Section 6.3), and it also makes the proof below more general.

We now show that $d_f$ satisfies the *triangle inequality*, and therefore is a metric.

**Theorem 1.** *For all $X, Y, Z$ with $S(X), S(Y), S(Z) \subseteq \{1, 2, \ldots, n\}$ we have:*

$$d_f(X,Y) \leq d_f(X,Z) + d_f(Z,Y)$$

*Proof.* We may assume that not both $X$ and $Y$ are $\emptyset$. If $d_f(X,Z)+d_f(Z,Y) \geq M$ we are done, since $d_f(X,Y) \leq M$. So we may assume that $d_f(X,Z)+d_f(Z,Y) < M$. Now

$$\begin{aligned} d_f(X,Y) &= \frac{\sum_{i=1}^{n} f(x_i,y_i)}{|S(X) \cup S(Y)|} = \frac{\sum_{i \in S(X) \cup S(Y)} f(x_i,y_i)}{|S(X) \cup S(Y)|} \\ &\leq \frac{\sum_{i \in S(X) \cup S(Y)} f(x_i,z_i) + \sum_{i \in S(X) \cup S(Y)} f(z_i,y_i)}{|S(X) \cup S(Y)|} \\ &= \frac{\sum_{i \in S(X) \cup T} f(x_i,z_i) + \sum_{i \in S(Y) \cup T} f(z_i,y_i)}{|S(X) \cup S(Y)|} \end{aligned}$$

where the set $T$ is defined by $T = S(Z) \cap (S(X) \cup S(Y))$. We have

$$\begin{aligned} \sum_{i \in S(X) \cup T} f(x_i,z_i) &= \sum_{i \in S(X) \cup S(Z)} f(x_i,z_i) - \sum_{i \in S(Z) \setminus T} f(0,z_i) \\ &\leq \sum_{i \in S(X) \cup S(Z)} f(x_i,z_i) - \frac{tM}{2} \end{aligned}$$

with $t = |S(Z) \setminus T|$. We conclude

$$\begin{aligned} d_f(X,Y) &\leq \frac{\sum_{i \in S(X) \cup S(Z)} f(x_i,z_i) + \sum_{i \in S(Y) \cup S(Z)} f(z_i,y_i) - tM}{|S(X) \cup S(Y)|} \\ &= \frac{d_f(X,Z)|S(X) \cup S(Z)| + d_f(Z,Y)|S(Y) \cup S(Z)| - tM}{|S(X) \cup S(Y)|} \end{aligned}$$

Now $-tM \leq -t(d_f(X,Z)+d_f(Z,Y))$ (because of the assumption that $d_f(X,Z)+d_f(Z,Y) < M$). So, noting that $|S(X) \cup S(Z)| = t + |S(X) \cup T|$ (and similarly for $|S(Y) \cup S(Z)|$) we get

$$\begin{aligned} d_f(X,Y) &\leq \frac{d_f(X,Z)|S(X) \cup T| + d_f(Z,Y)|S(Y) \cup T|}{|S(X) \cup S(Y)|} \\ &\leq d_f(X,Z) + d_f(Z,Y) \end{aligned}$$

since $|S(X) \cup T| \leq |S(X) \cup S(Y)|$ (and similarly for $|S(Y) \cup T|$). □ □

Before studying several properties of the metric, we first notice that its behaviour deviates from that of standard distance measures. As an example, if we have two given points, and we move one of these in a "new" dimension, the distance changes considerably whereas in the Euclidean case it does not.

Interesting properties of this measure are:

- If $X$ and $Y$ are "normal" sets, i.e., $x_i, y_i \in \{0, 1\}$ $(i = 1, 2, \ldots, n)$, we note that

$$d_f(X, Y) = f(1, 0)\frac{|X \setminus Y| + |Y \setminus X|}{|X \cup Y|} = f(1, 0)\left(1 - \frac{|X \cap Y|}{|X \cup Y|}\right)$$

- $d_f(\emptyset, (\underbrace{1, \ldots, 1}_{n})) = nf(1, 0)/n = f(1, 0)$.

Here we use the notation $(x_1, x_2, \ldots, x_n)$ for the multiset $X$, where again $x_i$ denotes the number of times the element $i$ occurs in $X$ (cf. the example in Section 6.2).

A variant of this measure can be defined as follows:

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

By using this measure, we can drop the separate definition of $\widetilde{d}_f(\emptyset, \emptyset)$. Another advantage of this measure is that $d(\emptyset, \{x\}) = d(\emptyset, \{x, y\}) = \ldots = \frac{1}{2}f(1, 0)$, while $\widetilde{d}(\emptyset, \{x\}) = \frac{1}{2}f(1, 0)$, $\widetilde{d}(\emptyset, \{x, y\}) = \frac{2}{3}f(1, 0)$ and so on. All conditions for a distance measure hold, since this function is still symmetric, the distance between identical multisets is zero, and the triangle inequality holds. To show the latter property, we can use a proof that is analogous to the one above, except for the last step, in which we replace $|S(X) \cup T|/|S(X) \cup S(Y)| \leq 1$ by $|S(X) \cup T|/(|S(X) \cup S(Y)| + 1) \leq 1$. Another way of proving it is by adding a new element $*$ that is present once in each multiset. This reduces the problem to the property shown above: $\widetilde{d}_f(X, Y) = d_f(X \cup \{*\}, Y \cup \{*\})$.

The application of weights for certain elements can be done by multiplying the number of elements to which the weight must by applied by the weight. These weights need not be integers, which is the reason why $f$ is defined on real numbers in Section 6.3. As an example, suppose we have the multiset $X = (1, 2, 1)$ and we want to apply the weight 10 to the first element. The resulting multiset $X'$ is defined by $X' = (10, 2, 1)$. We shall return to this issue in Section 6.4.

In order to obtain more reasonable and intuitive measures, the following restriction can be posed upon $f$:

$$f(x, y) \leq f(x', y') \quad \text{if } x' \leq x \leq y \leq y' \tag{6.6}$$

It then follows that $\lim_{k\to\infty} f(k, 0) = M$. In this case it is easy to show that the condition that $f(x, 0) \geq M/2$ is mandatory for the triangle inequality to

hold. Indeed, let $X = (k, 0, \ldots, 0)$, $Y = (0, \ldots, 0)$ and $Z = (0, \ell, \ldots, \ell)$. With $|S(X)| = 1$, $|S(Y)| = 0$ and $|S(Z)| = n - 1$, we have

$$\begin{aligned} d_f(X,Y) &= f(k,0) \to M \text{ when } k \to \infty \\ d_f(X,Z) &= \frac{f(k,0) + (n-1)f(\ell,0)}{n} \to \frac{M + (n-1)f(\ell,0)}{n} \text{ when } k \to \infty \\ d_f(Z,Y) &= \frac{(n-1)f(\ell,0)}{n-1} = f(\ell,0) \end{aligned}$$

Now $d_f(X,Y) \leq d_f(X,Z) + d_f(Z,Y)$ implies $f(k,0) \leq 2f(\ell,0)$ (let $n \to \infty$). With $f(\ell,0) < M/2$ for some $\ell > 0$ this is not true, so the triangle inequality does not hold.

A natural way to generate a suitable $f$ is the following. Start with a function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, and put $f(x,y) = |g(x) - g(y)|$. Clearly, properties (1), (2) and (4) hold for $f$. We may take $g(0) = 0$. If in addition $g$ is an increasing function with $\lim_{x\to\infty} g(x) = M$ and $g(x) \geq M/2$ for $x \in \mathbb{R}_{>0}$, $f$ also satisfies properties (3) and (6.6). If $g$ is injective, e.g., if $g$ is strictly increasing, (5) holds too.

Typical examples include:

- $g(x) = 1$ for $x$ with $0 < x \leq 1$ and $g(x) = M = 2$ for $x$ with $x \geq 1$
- $g(x) = 1/2$ for $x$ with $0 < x < L$ and $g(x) = M = 1$ for $x$ with $x \geq L$; here $L$ is a (large) constant
- $g(x) = 1/2$ for $x$ with $0 < x \leq 1$ and $g(x) = x/(x+1)$ for $x$ with $x > 1$ ($M = 1$), see Section 6.4; note that if we only use integer arguments, we just need the "$x/(x+1)$ part"
- $g(x) = 1/2$ for $x$ with $0 < x \leq 1$ and $g(x) = (2^x - 1)/2^x$ for $x$ with $x > 1$ ($M = 1$)

We conclude with a more intuitive explanation of the metric. Consider two vases filled with marbles of different colours. We first take a look at the marbles of the first colour. If both vases contain many marbles of this colour, the difference should be small, but the difference between one marble and no marbles should be large. The exact difference can be tuned by altering the function $f$, which specifies the distance between groups with a different number of marbles of the same colour.

When looking at all colours, we repeat the procedure above and divide by the amount of colours we have encountered. This differs from division by the total number of marbles, or by (some variation of) the total number of possible colours. This latter option, for instance chosen in case of the Euclidean distance, does not keep track of the "sizes" of the multisets under consideration. Choosing the total number of marbles as denominator — as the Bray-Curtis distance does — has the disadvantage that adding one marble of a fresh colour is hardly noticed, while our metric is much more sensible to this. Our metric emphasizes the number of different colours.

### Relation with other distance measures

Many well-known distance measures are special cases of the one we describe here. For example the Jaccard distance can be constructed by any $f$ with $f(1,0) = 1$, where the multisets must be "normal" sets, so $A = S(A)$ and $B = S(B)$. As noted before, this results in the following formula for sets:

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

To produce the Canberra distance (with the extended denominator) we use the following $f$:

$$f(x,y) = \frac{|x-y|}{x+y} \text{ for } (x,y) \neq (0,0)$$

and $f(0,0) = 0$. Note that this $f$ cannot be constructed by a function $g$ in the way explained above.

## 6.4 Applications

In this section we use the following function for $f$:

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

This function satisfies properties (1)–(6) mentioned in the previous section; it is a result of using $g(x) = x/(x+1)$. This function has the interesting property that if both $x$ and $y$ are large, the resulting value is small. For example, the (pairwise) distance between 0 and 1 is larger than the distance between 8 and 9, which is intuitive in many applications concerning multisets.

The visualisation algorithm we use in this section is a randomized push-and-pull oriented algorithm [43], comparable to a competitive neural network. It gives a projection of the original points in a 2-dimensional space. The reason we use this algorithm is because it is fast and able to give a clustering for many data points, where normal dimension reduction algorithms perhaps would fail. For the purpose of this chapter, there is no need to go into detail concerning this algorithm. We only state here that the Euclidean distances between points in the 2-dimensional space approximate the original distances as good as possible.

**Plagiarism:** When comparing two documents while ignoring the context and the semantics, we can make a multiset of words in the documents. To accommodate for the difference in lengths of two documents, we can increase the weight for each word in the smallest document by the relative size of the documents. In this way, identical copies of the same text will be detected.

In this chapter we will not further elaborate on this; we only mention the flexibility of the measures, which allows for many user-defined alterations.

**Genomics:** We will give an example of an instance of our distance measure in the genomics domain. We do not claim this particular instance is the best for clustering species, but from the illustrative example it can be seen that this instance does work.

A *genome* of some biological species can be considered as a long string over a small finite alphabet, usually $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$; it can be converted into a multiset by using a sliding window of length $n$ to count the occurrence of each substring (or factor) of length $n$. Of course the number of occurrences will depend on $n$: the larger $n$ is, the lower the number of occurrences will be on average.

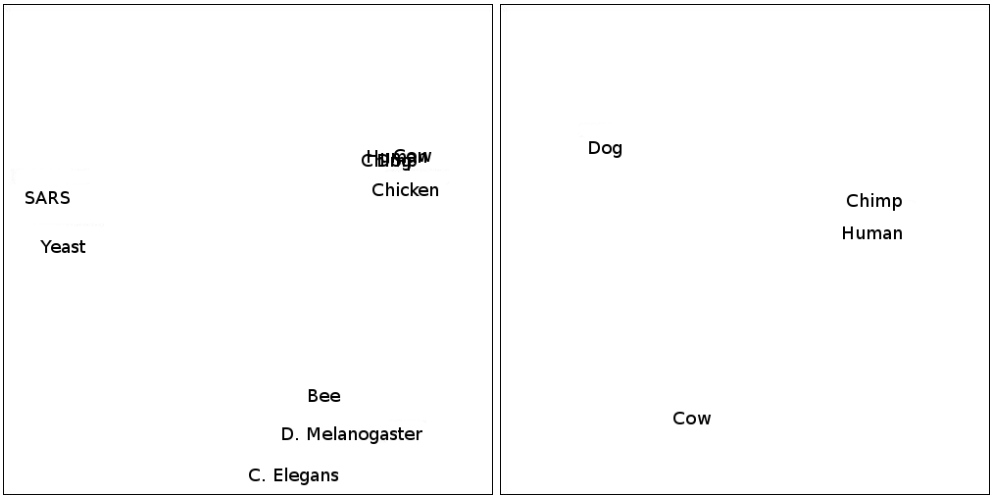


Figure 6.1: Visualisation for ten species; left: all ten; right: the four mammals

By determining the number of occurrences of each factor in two genomes, we obtain two multisets that can be compared to each other. If we use our distance measure with the function mentioned above, the occurrences of unique or almost unique substrings will account for most of the difference between the genomes. Factors that occur many times in both genomes are accounted for accordingly.

In this way we compare two species mostly on the number of differences between rare substrings in their DNA. In Figure 6.1 a clustering based on this distance is shown; DNA [69] of ten species (SARS, Yeast, Bee, C. Elegans, Drosophila Melanogaster, Chicken, Cow, Dog, Chimp and Human) is used; the right part of the figure zooms in on the four mammals, which are very close together in the left part of the figure (the labels are practically on top of each other). The sizes of the genomes vary from $3.69 \cdot 10^4$ for SARS to $3.60 \cdot 10^9$ for Cow. As in the case of Plagiarism, we here also compensate for the difference in sizes.

Apart from this type of clustering, other visualisations are possible too: the metric can also be used to generate a phylogenetic tree, for example.

## Criminal records

For comparing criminal records [9], the above function is very well suited. When we make a multiset from criminal records, we get for example a multiset where the first element represents bicycle theft, the second one represents violent crimes, and so on. The difference between no crime and one or more crimes in each category accounts for a large difference, while having two large numbers

in each category accounts for almost no difference at all. This is rather useful, since two people who steal bikes on a regular basis, can be seen as much alike.

Of course there are some differences between the categories which one might want to accentuate. For example, a murder is considered a much more severe offence than a bicycle theft. One way to accommodate for this difference is to use a vector of weights $W = (w_1, w_2, \ldots, w_n)$ with $w_i \in \mathbb{R}_{\geq 0}$ and to make the following adjustments to the distance measure:

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

It is easy to prove that this adjustment does not change the fact that the distance formula is still a good metric.

Now, by choosing the vector of weights carefully (this must be done by an expert in criminology) we can assign relative weights for crimes. In our example, we can set the weight for bicycle theft to 1 and the weight for murder to a large integer to accentuate the severity of the crime.

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. 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 all-round criminal.

| | 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 6.1: Ten criminals, four crimes

In the top-left picture of Figure 6.2 we see a clustering of these ten criminals with the standard $f_0$. In the picture right next to it, we applied weights $1, 10, 100, 1000$, respectively, to the crimes, to specify the weight of the crime. 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.

In the bottom-left picture, we see a clustering with $f$ chosen in such a way that we get the Jaccard distance, so we treat the criminals as sets. Notice that criminals 2 and 3 now have distance zero to each other (the labels are on top of each other in this picture). The bottom-right clustering uses a totally different $f$: $f_1(x, y) = \frac{3}{2} - f_0(x, y)$ for $(x, y) \neq (0, 0)$ and $f_1(0, 0) = 0$. Note that, e.g., $f_1(0, 1) = 1 > \frac{3}{4} = f_1(0, 3)$, so property (6.6) does not hold for $f_1$. Now criminals with disjoint behaviour are grouped, leading to a "dissimilarity" clustering.

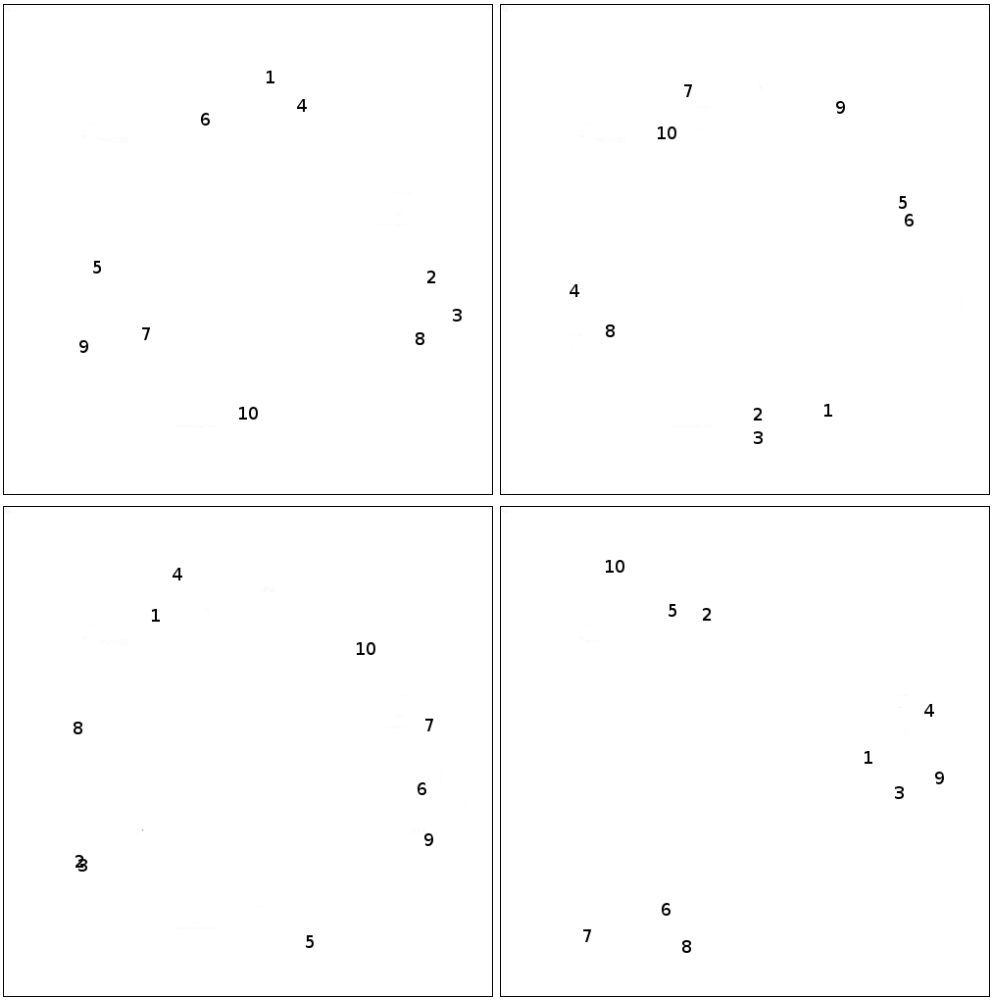


Figure 6.2: Four different clusterings for ten criminals

## 6.5 Conclusions and further research

In this chapter we have proposed a new flexible distance measure, that is suitable in many fields of interest. It can be fine tuned to a large extent.

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. We also want to apply the measure to a real, large database. Finally, we would like to examine the relation with more statistically oriented measures.